

Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN
SCIENZE BIOTECNOLOGICHE E FARMACEUTICHE

Ciclo XXXI

Settore Concorsuale: 03/D1

Settore Scientifico Disciplinare: CHIM/08

FREE ENERGY AND KINETICS IN PROTEIN-LIGAND BINDING:
EXPERIMENTAL MEASUREMENTS AND COMPUTATIONAL ESTIMATES

Presentata da: Dorothea Gobbo

Coordinatore Dottorato

Prof. Santi Mario Spampinato

Supervisore

Prof. Andrea Cavalli

Co-supervisori

Prof. Pietro Ballone
Dr. Sergio Decherchi

Esame finale anno 2019

Contents

Preface	4
1. Introduction	5
1.1. Thermodynamics and binding affinity.....	7
1.2. Drug-target kinetics in drug discovery	9
1.3. The most popular biophysical experimental techniques to characterize PL interactions	12
2. Theoretical background	15
2.1. Statistical mechanics.....	15
2.1.1. The canonical ensemble.....	16
2.1.2. The grand-canonical ensemble	19
2.1.3. The ideal free energy term.....	20
2.1.4. Molecular systems	22
2.2. Potential energy surface (PES)	25
2.2.1. The Born-Oppenheimer approximation.....	26
2.2.2. Stationary points of the PES	27
2.2.3. The force field	29
2.2.3.1. The Amber force field	32
2.2.4. Density Functional Theory (DFT)	33
2.3. The harmonic and quasi-harmonic approximation for molecular vibrations.....	35
2.4. Molecular dynamics (MD)	40
2.5. Monte Carlo (MC) methods	43
2.6. Rare events	45
2.6.1. Transition state theory (TST)	46
2.7. Free energy calculations	47
2.7.1. Free energy perturbation theory.....	49
2.7.2. Thermodynamic integration	50
2.7.3. Relative free energy determination.....	51
2.8. Enhanced sampling methods in drug discovery	52
2.9. Methods to compute the absolute free energy: an overview	56
3. Understanding protein-ligand unbinding kinetics in kinases through electrostatics-driven adiabatic bias molecular dynamics.....	62
3.1. Aim of the project.....	62

3.2.	Computational methods.....	64
3.2.1.	Electrostatics-driven adiabatic bias molecular dynamics (elABMD).....	64
3.2.1.1.	Adiabatic bias molecular dynamics (ABMD)	64
3.2.1.2.	Electrostatics-driven collective variable (eCV).....	65
3.2.2.	Simulation Setup and Analysis.....	66
3.3.	Experimental methods	68
3.3.1.	Chemistry	68
3.3.2.	GSK-3 β expression and purification	71
3.3.3.	Kinetic characterization of GSK-3 β inhibitors by Surface Plasmon Resonance (SPR)	71
3.3.4.	Analysis of binding data	72
3.3.5.	Crystallization of GSK-3 β in complex with compounds 4-6	73
3.3.6.	Data collection and structure determination	74
3.4.	Results	75
3.4.1.	A retrospective validation of elABMD protocol: the GK case.....	75
3.4.1.1.	Validation of elABMD protocol.....	75
3.4.1.2.	Unbinding path analysis and Structure-Kinetic Relationships (SKRs)	78
3.4.2.	A prospective application of elABMD protocol: the GSK-3 β case.....	85
3.4.2.1.	Prospective predictions.....	85
3.4.2.2.	Explanation of protein-ligand unbinding paths	87
3.5.	Discussion and conclusions	91
4.	A computational approach to estimate absolute free energies and hydration free energies in atomistic simulations.....	93
4.1.	Aim of the project.....	93
4.2.	Flow diagram of our free energy computation	93
4.3.	Simulation setup	95
4.3.1.	The models	95
4.3.2.	Units	95
4.3.3.	System preparation	96
4.3.4.	Computation and diagonalization of the Hessian matrix.....	97
4.3.5.	Volume optimization by quasi-harmonic (QH) approximation.....	98
4.3.6.	Thermodynamic perturbation	102
4.3.7.	Free energy decomposition.....	105

4.3.7.1.	Quasi-harmonic (QH) contribution.....	105
4.3.7.2.	Ideal contribution.....	106
4.3.7.3.	An-harmonic contribution from perturbation theory	107
4.4.	Results	107
4.4.1.	Validation of the reference system	108
4.4.2.	Recovering the full an-harmonicity	114
4.4.2.1.	Addressing the molecular diffusion of fluid samples	115
4.4.3.	First test application: Hydration free energy (HFE) estimates	123
4.4.3.1.	Detailed analysis of ketoprofen in the equilibrium crystal phase	132
4.5.	Scaling to large systems	139
4.6.	Discussion and conclusions	151
5.	Conclusions	154
6.	Acknowledgments	157
7.	Appendix	158
7.1.	Median unbinding time based-ranking correlations of GK series	158
7.2.	Median unbinding time based-ranking correlations of GSK-3 β series.....	160
7.3.	Statistics of randomly selected 10 production runs of GSK-3 β inhibitors unbinding simulations	161
7.4.	Compounds 5 and 7	163
7.5.	Surface plasmon resonance (SPR) affinity and binding curves.....	165
7.6.	Further test of the force field	171
7.6.1.	Methane	171
7.6.2.	Propionic acid.....	172
7.6.3.	Piperidine.....	173
7.6.4.	Nitromethane	173
7.7.	Validation of the quasi-harmonic (QH) equilibrium volume	175
8.	Bibliography	177
	Abstract	191

Preface

Molecular organization and recognition drive all biological processes, which are thus dependent on how macromolecules interact with each other. Thermodynamics and kinetics largely determine the dynamical processes underlying how biomolecules behave *in vivo*. Therefore, an accurate characterization of the energy and free energy aspects governing the formation of supramolecular complexes is a crucial pre-requisite for a deep understanding of molecular principles driving biological interactions.

This topic is particularly relevant in drug design, which aims at understanding drug action at the molecular level to guide the rational design of new medicines. A therapeutically relevant drug response requires the availability of the drug molecule for binding to the biological target and the translation of the interaction to a selective physiologic response. The binding process involves the de-solvation of the small molecule, the approach of the drug to the pharmacological target, followed by the establishment of specific non-covalent interactions. The stability and the duration of the drug-target complex contribute to the pharmacological response. In this context, thermodynamics provides the driving force and kinetics describes the rates of transitions between energy basins. Thermodynamics of protein-ligand binding is quantified by the binding free energies, ΔG_{bind} , or equilibrium dissociation constants, K_d , which informs on the affinity of a ligand under equilibrium conditions. Since drug and target are out of equilibrium *in vivo*, the thermodynamic description of protein-ligand binding needs to be complemented by the knowledge of kinetic association and dissociation rates, k_{on} and k_{off} . Experimental biophysical techniques, such as isothermal titration calorimetry (ITC) and surface plasmon resonance (SPR), are available to characterize the thermodynamics and kinetics of protein-ligand binding. Conversely, the computational counterpart able to efficiently predict thermodynamic and kinetic properties still faces severe challenges mainly due to the limited force field accuracy and the high computational costs. Indeed, the prediction of these properties requires extensive sampling of the conformational space characterized by free energy barriers leading to dissociation times far longer than the time scales usually sampled by computer simulations.

In this framework, my PhD program has been focused on addressing both thermodynamics and kinetics of protein-ligand complexes by computational approaches. First, I worked on the development of a new computational protocol to prioritize series of compounds on unbinding kinetics (Chapter 3). Then, I worked on a new method to estimate free energies that would be applicable to systems of arbitrary complexity (Chapter 4). Chapter 1 introduces the concepts of thermodynamics and kinetics with a particular focus on drug discovery. In Chapter 2, the theoretical background on simulation approaches applied to characterize protein-ligand binding interactions and to predict thermodynamic and kinetic properties is discussed. Finally, some general conclusions with a focus on the open challenges are given in Chapter 5.

1. Introduction

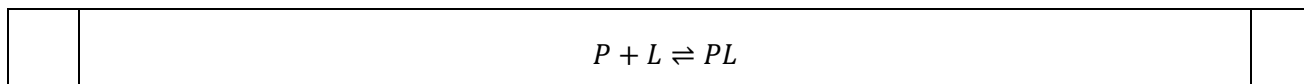
Virtually all biochemical activities of cell physiology, such as enzymatic reactions, signal transduction, and gene transcription, are mediated by the formation of transient binary complexes involving biological macromolecules.¹ Specificity and affinity are the main features of an efficient molecular recognition, which acts as a pre-requisite for all processes requiring interacting partners.²

Among biological macromolecules, proteins represent a very important class playing crucial roles in a huge number of cellular processes. The onset of direct physical interactions between proteins and binding partners, such as other proteins or peptides, nucleic acids, membranes, substrates, and small molecules, is required to their biological functions.³ In the following, the term “ligand” refers to any molecule able to bind a protein.

A complete comprehension of protein functions at the atomic level relies on the knowledge of their structural three-dimensional geometries and of the mechanisms driving the molecular recognition. Nowadays, essential steps of the biophysical characterization of biochemical processes are proteins’ expression, purification, and crystallization. The growing datasets of structural information collected by X-ray diffraction and NMR spectroscopy can give structural insights into the specific interactions characterizing protein-ligand complexes.⁴

Three different models have been proposed to describe protein-ligand binding mechanisms. In 1894, Emil Fischer introduced the “lock-and-key” model,⁵ which associates rigid conformations to the interacting partners. In the framework of the “lock-and-key” model, protein and ligand interact only when their binding interfaces are structurally complementary. This model is inadequate to explain experimental evidences stating that unbound conformations of binding partners can differ from the bound ones. Then, the “induced fit” model⁶ was proposed. It assumes that the protein binding site is flexible and that the interacting partner can induce a conformational change upon binding. Therefore, proteins have single native conformations in solution and are involved in minor conformational changes only upon interacting with a substrate. The inherent flexibility and dynamics of protein structures are taken into account in the “conformational selection” model,⁷ which sees proteins as an ensemble of conformational states coexisting in equilibrium with different population distributions. Therefore, binding partners can bind selectively to the most favorable conformational state, inducing a population shift and a redistribution of the available states. Since aspects of all three models have been observed experimentally, it is important to note that they can act simultaneously or in a sequential manner during binding events. More detailed discussions about this argument are reviewed elsewhere.⁸⁻¹¹

Let us assume a two-state binding process, in which protein, P , and ligand, L , associate by non-covalent interactions to form the binary protein-ligand complex, PL .



Protein-ligand interactions occur by multiple-states mechanisms, whose apparent rate constants consist of multiple elementary rate constants describing the transition between unbound, intermediate, and bound states.¹² For the sake of simplicity, the following discussion will consider this simplistic two-state mechanism comprising a single elementary step without any intermediate states.

The binary binding process is characterized by the equilibrium association constant, K_a , expressed as the ratio between the concentration of the complex, PL , and of the dissociated interacting partners, P and L . The binding affinity is quantified by the reciprocal of the association constant (i.e. the equilibrium dissociation constant, K_d) corresponding to the ligand concentration at equilibrium for which an equal probability of bound and unbound protein is achieved.

	$K_d = \frac{[P][L]}{[PL]}$	(1.1)
--	-----------------------------	-------

At equilibrium, K_d is directly related to the free energy difference between bound and unbound states:

	$\Delta G_{binding}^\circ = k_B T \ln \left(\frac{K_d}{C^\circ} \right)$	(1.2)
--	---	-------

where C° is a constant defining the standard concentration.

In contrast to K_d , which is directly related to the stable inter-molecular interactions between ligand, protein, and solvent, the rate constants, k_{on} and k_{off} , depend on transient interactions along the interactions pathways. Specifically, the rate constants are related to the highest free energy barrier (i.e. the transition state) separating the bound and unbound states:

	$k_{on} \propto e^{-\Delta G_{on}^\ddagger / RT}$	$k_{off} \propto e^{-\Delta G_{off}^\ddagger / RT}$	(1.3)
--	---	---	-------

where ΔG_{on}^\ddagger is the free energy difference between the dissociated states, P and L , and the transition state, and ΔG_{off}^\ddagger is the free energy difference between the binary complex and the transition state.

Thermodynamics and kinetics of binding are linked by the following relationship between the equilibrium dissociation constants and the kinetic rates:

	$K_d = \frac{[P][L]}{[PL]} = \frac{k_{off}}{k_{on}}$	(1.4)
--	--	-------

This relationship is particularly important as it highlights how the difference in free energy between the bound and unbound state, which is a direct measure of the binding affinity, is related to the free energy of the transition state, ΔG^\ddagger .

In the next paragraph, the basic thermodynamic concepts and relationships are introduced. Then, the role of kinetics in drug discovery is discussed.

1.1. Thermodynamics and binding affinity

A protein-ligand-solvent system is a thermodynamic system composed of the solute (i.e. protein and ligand molecules) and the solvent (i.e. bulk water). In such a system, there are very complex interactions and heat exchange among these substances. The laws of thermodynamics dictate the relationships between the concentrations of the binding partners and how the heat transfer is related to the various energy changes.

The forces driving protein-ligand binding comprehensively include different interactions and energy exchanges among solutes and solvent. Gibbs free energy, which is defined as the thermodynamic potential measuring the capacity of the system to do maximum or reversible work at a constant temperature and pressure (isothermal, isobaric), is one of the most important thermodynamic quantities for the characterization of binding forces.¹³⁻¹⁴

In analogy with any spontaneous process, protein-ligand binding occurs only when the change in Gibbs free energy, ΔG , is negative at constant temperature and pressure and equilibrium conditions. Because the protein-ligand association extent is determined by the magnitude of the negative ΔG , it can be considered as a measure of the stability of the binary complex or of the binding affinity between the interacting partners.

The binding free energy can be defined as a function of the equilibrium dissociation constant, as stated by Equation 1.2. The difference in free energy can also be parsed into its enthalpic and entropic contributions as:

	$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$	(1.5)
--	---	-------

where ΔH° and ΔS° are the change in standard enthalpy and standard entropy upon ligand binding, respectively, and T is the temperature in K.

Enthalpy is a measure of the total energy of a thermodynamic system. ΔH° is negative in exothermic processes that occur when energetically favorable non-covalent interactions are established between protein and ligand. ΔH° is positive in endothermic processes in which energetically favorable non-covalent interactions are disrupted. For a binding process, ΔH° reflects the energy change upon binding to the protein,

and it is usually treated as the difference in energy resulting from the formation of non-covalent interactions at the binding interface. However, the ΔH° associated to a binding reaction is a global property including not only contributions from the interacting solutes, but also from the solvent. Indeed, the change in enthalpy upon binding is the result of forming and breaking many individual non-covalent interactions, such as those established between the protein/ligand and solvent and those due to the reorganization of the solvent molecules at the complex's interface. These individual components may make favorable and unfavorable contributions. The net enthalpy difference is the result of the combination of these terms.

Entropy is a measure of how the heat change will be distributed over the entire thermodynamic system. The second law of thermodynamics states that the heat always flows spontaneously from regions of higher temperature to regions of lower temperature. In this process, the initial order of the system decreases, and the entropy is seen as a measure of the degree of disorder or randomness of the system. Entropy is a state of function, and ΔS° is a global thermodynamic property. A positive change in entropy indicates the overall increase of the system's disorder, whereas a negative ΔS° indicates the decrease of degree of freedom of the system.

The total entropy change associated with binding may be parsed into three terms:

	$\Delta S^\circ = \Delta S_{solv}^\circ + \Delta S_{conf}^\circ + \Delta S_{rt}^\circ$	(1.6)
--	--	-------

ΔS_{solv}° represents the change of entropy due to the solvent. Upon binding, the solvent molecules at the interacting interface are released, making a favorable contribution to the binding entropy. ΔS_{conf}° reflects the change in the conformational freedom of both protein and ligand upon binding. ΔS_{conf}° can give positive or negative contributions to the binding entropy depending on the increased or reduced degrees of freedom of the complex in comparison to the unbound states. ΔS_{rt}° represents the loss of translational and rotational degrees of freedom of protein and ligand upon binding and contributes unfavorably to the binding entropy. Then, binding reactions are likely when the entropic penalties due to ΔS_{rt}° are overcome due to large positive energetic contributions from ΔS_{solv}° or ΔH° .

A spontaneous binding event occurs when the change of the system free energy is negative. The extent of the difference in free energy between unbound and bound states determines the stability of the complex. Therefore, one can assume that the decrease of the system free energy drives the protein-ligand binding. Considering that the sign and magnitude of the free energy is determined by the change in enthalpy and entropy, those are the driving factors that contribute to the protein-ligand binding. Moreover, enthalpy and entropy are, to some extent, competing quantities, whose changes upon binding can compensate each other resulting in small or absent enhanced binding affinity (enthalpy-entropy compensation). This phenomenon

may be rooted in the formations and disruptions of weak non-covalent interactions in the thermodynamic system.

Because of this compensation, the discrimination between the entropic and enthalpic contributions to the binding free energy is fundamental, particularly in drug discovery. Indeed, the thermodynamic approach to optimize protein-ligand binding affinity requires the knowledge and understanding of how to alter the structure of the small molecule in order to gain favorable contributions in terms of binding enthalpy and entropy. Modifying the structure of the ligand to enhance beneficial change in its binding enthalpy involves optimizing existing interactions with the protein and/or increasing the number of non-covalent interactions.

These efforts may be driven by experimental crystal structures of protein-ligand complexes, despite the intrinsic limitations associated to these structures, such as the uncertainty in atomic positions and the absence of information regarding disordered water molecules. The rational design of strong non-covalent interactions between polar groups of protein and ligands is fairly difficult in practice because the energy associated to those interactions is distance and angle-dependent. Another factor that complicates the enthalpy optimization is that small molecules rarely bind in the conformation corresponding to a global energy minimum.

The binary complex formation is associated to a loss in rotational and translational entropy. Strategies to overcome this problem rely on introducing a conformational constraint, usually a ring, into the molecule to stabilize its biological active conformation in solution. Another strategy requires the addition of a non-polar group to the ligand structure in order to enhance the change in entropy from desolvating non-polar surfaces.

Thus, this ligand-centric approach to enhance binding affinity is mainly focused on the non-bonded interactions between protein and ligand and on the conformational and desolvation factors associated with the small molecule itself. However, in this way, the contributions due to the conformational changes in the protein structure upon binding are ignored.¹⁵⁻¹⁶

While high affinity for a target is the basic requirement for any potential drug candidate, thermodynamics only is not enough to comprehensively characterize protein-ligand binding and to fully account for time-dependent changes of solutes' concentrations at *in vivo* conditions. Moreover, optimizing the binding affinity, the energy of the ground state associated to the binary complex is affected with unpredictable effect on the kinetics of binding, which depends on the height of the highest energy barrier, (i.e. the transition state). Therefore, both thermodynamics and kinetics of molecular systems have to be taken into account.

1.2. Drug-target kinetics in drug discovery

For binding to occur, the initial contacts/collisions between protein and ligand have to form an encounter complex. In this context, molecular diffusion plays a decisive role.¹⁷ Diffusion originates from molecular

kinetic energy and is the entropy-driven process governing the binary complex formation. In the thermodynamic system including solutes (protein and ligand) and solvent, the diffusion of solute molecules originates from the kinetic energy of the solute themselves, as well as the collisions of the protein/ligand molecules with the water molecules, which move with different velocities in different random directions. Considering the high number of water molecules, the random collisions between solvent and solute molecules may play a role in facilitating the rotations and translations of solutes and their final interactions.¹⁸ Long-range electrostatic interactions promote the association of partners with opposite charges, overcoming the diffusion limit.¹⁹

The collision theory was the first attempt to provide an analytical description of the dependence of the rate constant of a reaction on the temperature and activation energy. The relation is expressed by means of the Arrhenius equation:²⁰

	$k = Ae^{-\frac{E_a}{RT}}$	(1.7)
--	----------------------------	-------

where E_a is the activation free energy of the underlying process, T is the temperature, R is the universal gas constant. The pre-exponential factor, A , also called frequency factor, is a constant that can be determined experimentally or numerically. It quantifies the number of times two molecules collide. Note that not every collision results in the expected product, since a number of factors, such as the orientation between the interacting partners, are required.

Within the collision theory,²¹ the pre-exponential factor associated to the collision of two particles, A and B, can be defined as:

	$A = d_{AB}^2 \sqrt{\frac{8k_B T}{\mu}}$	(1.8)
--	--	-------

where d_{AB} is the collision radius between particles A and B, k_B is the Boltzmann constant, and μ is the reduced mass of the system. The Arrhenius rate law has been widely used to determine the energies for the reaction barrier, ignoring any mechanistic considerations. However, the collision theory deals only with gases and does not account for structural complexities molecules and biomolecules in particular.

In order to resolve this discrepancy, the transition state theory (TST)²² was developed to give a more accurate representation of the pre-exponential factor yielding to the corresponding rate. TST is based on three fundamental concepts. First, the reaction rates can be studied by examining activated complexes that lie near the saddle point of the potential energy surface. Second, the activated complexes and the reactants are in the so-called, quasi-equilibrium state. And finally, the activated complexes can be converted into products and the rate of conversion can be computed by the kinetic theory. To accurately apply TST, a description of the

potential energy surface of the system is required, but the characterization of the complex structure (and of the target, in particular) at the transition state is not straightforward.

In the framework of a drug-target binding as a two-state process, the association and dissociation rates are controlled by the difference in free energy between the ground and the corresponding transition states (Eq. 1.3). The dissociation rate constant, k_{off} , can be seen as an indicator of the fit or complementarity of the compound to the target in the bound form. The tighter the interaction, the smaller the k_{off} will be. As such, k_{off} is independent of the target/ligand concentration, and it is expressed in s^{-1} . By contrast, the association constant, k_{on} , is a measure of the fit of compound to the target when both are still in the unbound form. If no wide conformational changes are required for binding to occur, the on-rate will be limited only by diffusion. The k_{on} is a second-order rate constant with the unit $M^{-1}s^{-1}$.

In the drug-discovery field, kinetics has received increasing attention in recent years,²³ following the perspective article of Copeland published in 2006.²⁴ His discussion starts from the evidence that a drug is efficacious when it is bound to its physiological target, whose cellular function is consequently somehow modulated.²⁵⁻²⁶ Of course, there are exceptions in which the modulation of the biological target can persist subsequently to drug dissociation. Accounting that the association of the drug and target in the binary complex is the precursor of the drug action *in vivo*, drug discovery programs have been focused on the optimization of target affinity and selectivity. Target affinity is experimentally assessed in cell-free assays measuring the binding of the substrate to a target directly or indirectly through the effect of the compound on the biological activity of the receptor. These measurements are performed *in vitro* under closed-system conditions, in which the protein is exposed to fixed concentrations of the binding partner. Then, target-binding affinity is quantified in terms of half-maximal inhibitory concentration, IC_{50} , or by the equilibrium dissociation constant, K_d , for the binary complex. In theoretical studies, binding affinity is computed through the calculation of the free energy difference between bound and unbound states.²⁷ There are cases in which equilibrium dissociation constant is directly related to the *in vivo* efficacy of the drug. However, more often only a qualitative relationship between those quantities is observed. Thus, *in vitro* experiments can only poorly reproduce the dynamic process of protein-ligand binding under the open-system conditions of *in vivo* settings. In this context, Copeland suggests that the duration of the pharmacological effect of a substrate *in vivo* depends on the stability of the binary complex, whereas the efficacy of a drug can be related to the kinetic rate of association and, more critically, on the dissociation rate constant of the binary complex.^{24, 28}

Therefore, the residence time defined as the reciprocal of the dissociation rate constant, $t_r = 1/k_{off}$, and the complex half-life, $t_{1/2} = (\ln 2)/k_{off}$, become key quantities to evaluate during the lead optimization.

Another fundamental property to optimize during drug discovery campaigns is the selectivity, which refers to the relative ability of a drug to engage the chosen target compared to off-target molecules, providing valuable insights into possible unwanted side effects. Selectivity can be determined from affinity-based

measurements by comparing the IC_{50} or K_d values determined against two different targets. This kind of selectivity is actually *thermodynamic* selectivity since both IC_{50} and K_d are measured at equilibrium conditions. Since a compound can have the same affinity for two proteins but different association and dissociation rates, affinity-based assessments of selectivity are not able to give insights into the *kinetic* selectivity of a ligand with respect to different targets. The contribution of kinetic selectivity to the therapeutic window is intimately related to the time-dependence of drug concentration at the target site (i.e. pharmacokinetics, PK). Drugs that eliminate rapidly relative to the lifetime of the drug-target complex will maximize the potential benefit of kinetic selectivity in situations where prolonged occupancy of the target is mitigated. Target turnover also impacts kinetic selectivity, since the rapid synthesis of new target will negate the effects of prolonged target occupancy at low drug concentration.²⁹

Although the binding kinetics and residence time complement the information regarding drug-target affinity giving insights into drug efficacy *in vivo*, the overall binding process is not that simple. The concentration profile of a drug is dictated by its PK properties, which determine if the residence time can have an impact on the duration of the pharmacological effect *in vivo*. It was demonstrated³⁰ that combining the pharmacokinetic and binding kinetic information, the prolongation of binding owing to a long drug-target residence time can occur when the binding dissociation is slower than the pharmacokinetic elimination. However, experimental data on commercial drugs reveal the opposite, suggesting that the evaluation of the drug-target residence time need to be complemented with additional considerations to estimate the duration of the therapeutic effect *in vivo*.³¹

1.3. The most popular biophysical experimental techniques to characterize PL interactions

The structural and dynamic data alone, even when coupled with computational methods, cannot provide information regarding the complete thermodynamic and kinetic profile consisting of binding free energy, enthalpy, entropy, and kinetic rates. Nowadays, the availability of molecular biology techniques to produce large amounts of purified proteins allows to routinely measuring affinity and kinetic constants *in vitro*, as well as to determining atomic-resolution crystal structures of proteins in their unbound and bound states.

Many experimental techniques might be used to investigate various aspects of protein-ligand binding. In the next paragraphs, isothermal titration calorimetry (ITC) and surface plasmon resonance (SPR) are briefly presented. For more detailed discussions about ITC and SPR, I refer to Ref. 32 and 33.

Isothermal titration calorimetry (ITC) is the only approach able to measure directly the heat exchange during the complex formation at constant temperature, becoming a gold standard in determining the energies driving the binding process and stabilizing the inter-molecular interactions.³²⁻³³

A typical ITC experiment requires three steps. First, known aliquots of ligand are titrated into a solution containing the biological target causing heat to be either released or absorbed. As a consequence, a temperature imbalance between the reference and sample cells is measured. Such an imbalance is compensated for by modulating the feedback power applied to the cell heater, which increases and decreases in endothermic and exothermic reactions, respectively. The overall measurements consist of the time-dependent input of the power required to maintain equal temperatures between the sample and reference cells at each titration. The primary ITC data are the power applied to the sample cell as a function of time. These data are processed to obtain the binding curve representing the heat of reaction per injection as a function of the ratio of the total ligand concentration to the protein concentration. Finally, fitting the binding curve, the binding constant K_b , the binding enthalpy ΔH , and the stoichiometry of the binding event n are obtained. Knowing the binding constant, the standard Gibbs binding free energy ΔG° , and the binding entropy ΔS° can be derived. Moreover, when ITC experiments are performed on a range of temperatures, the heat capacity at constant pressure, ΔC_p , can be obtained by determining ΔH° varying the system temperature. Because of the strong correlation between ΔC_p and the surface area buried on forming a complex, ΔC_p provides a link between thermodynamic parameters and the structural information of proteins. The hydrated water and bulk water will lead to a change in heat capacity proportional to the amount of surface area involved. The desolvation of both the protein and ligand upon binding can make positive or negative contributions to ΔC_p , depending on the burial of the polar or apolar surface areas, respectively.³

Note that the heat exchange detected by ITC is the total heat effect in the sample cell upon the ligand addition, including not only the heat absorbed or released during the binding reactions, but also the heat effects arising from the dilution of the ligand and protein, the mixing of two different solutions, the different temperatures between the sample and reference cells, and so on. Therefore, evaluating the heat change due to the contribution of binding only is not straightforward.

Surface plasmon resonance (SPR) spectroscopy³⁴ is one of the most popular techniques used for the determination of association and dissociation rate constants during protein-ligand (un)binding events. SPR is an optical-based method that measures the change in the refractive index near the sensor surface. It is a label-free technique, which is advantageous in comparison with the radioligand binding assays that have been previously used for the biochemical characterization of the formation of specific drug-target complexes. Moreover, SPR spectroscopy is capable of real-time quantification of protein-ligand binding kinetics and affinities.

In the most popular configuration, the sensor surface is a thin gold film on a glass support, which is positioned on the bottom of the flow cell through which an aqueous solution flows continuously. The receptor molecules are immobilized on the sensor surface and the small molecule (i.e. the analyte in the SPR formalism), is injected into the aqueous solution. As the analyte binds to the immobilized receptor, an increase in the refractive index is observed. Once all the binding sites are occupied, running buffer without

analyte is injected through the flow cell to let the ligand molecules dissociate from the target protein. As the analyte dissociates, a decrease in the refractive index is measured. The time-dependent resonance unit (RU) curve is processed and fitted to determine the association and dissociation rates, k_{on} and k_{off} , and the equilibrium dissociation constant as the rate between k_{off} and k_{on} . Moreover, the equilibrium dissociation constant, K_d , is quantified also by fitting the resonance unit sinusoidal curve as a function of the analyte concentration. From SPR measurements, highly reproducible affinity measurements can be provided. Moreover, the binding enthalpy can be estimated by van't Hoff analysis.³⁵ Note that by immobilizing the protein, the conformational, translational, and rotational entropies may be affected impacting on the evaluation of the association rate constant.

2. Theoretical background

Biological systems can be described by computer simulations through microscopic models in which molecules are represented by interacting particles. The correspondence can be atomistically detailed, or especially for large molecules, it can be coarse-grained with particles representing groups of atoms.³⁶

Since molecules often consist of large assemblies of atoms, statistical mechanics provides the most suitable description of their behavior. Statistical mechanics, in turn, could be developed in the quantum mechanics (QM) framework, or in the classical mechanics (CM) one. Although reality is quantum mechanical, biophysical events occur at conditions that can be described by classical mechanics to a fair degree of accuracy. More importantly, while classical mechanics methods are well developed and widely used to study large systems, quantum mechanics is inherently more difficult or at least much more expensive to be used as a standard simulation approach.

2.1. Statistical mechanics

Statistical mechanics³⁷ aims to study the macroscopic properties of a system made by many particles starting from their microscopic descriptions.

Classical mechanics represents a generic molecular system as a set of coordinates $\{\mathbf{r}_i, i = 1, \dots, N\} \equiv \{\mathbf{r}^N\}$ and conjugate momenta $\{\mathbf{p}_i, i = 1, \dots, N\} \equiv \{\mathbf{p}^N\}$, where \mathbf{r}_i is the three dimensional vector identifying the position of particle i in Cartesian space. Degrees of freedom consist of coordinates and momenta, spanning the system phase space. Together, they define a point of the phase space, $\Gamma(\mathbf{r}_i, \mathbf{p}_i)$, namely a microstate.

The evolution of the system is described by a trajectory, $\{\Gamma(t)\} = \{\mathbf{r}_i(t), \mathbf{p}_i(t)\}$, connecting the points in the phase space visited over time. The definition of the system's total energy relies on the Hamiltonian function (an operator in QM), which is defined as the sum of kinetic and potential energy:

	$\hat{H} = \hat{T} + U = \sum_i \frac{\mathbf{p}_i^2}{2M_i} + U(\{\mathbf{R}_i\})$	(2.1)
--	--	-------

where M_i is the mass of the particle i .

In statistical mechanics, each macroscopic property, which is time-independent at equilibrium conditions, is the average of a microscopic operator $O(\mathbf{r}_i, \mathbf{p}_i)$ over all accessible states $(\mathbf{r}_i, \mathbf{p}_i)$, weighted by the probability $\rho(\mathbf{r}_i, \mathbf{p}_i)$ of visiting each state:

	$\bar{O} = \langle O \rangle = \int_{\mathbf{r}_i, \mathbf{p}_i} O(\mathbf{r}_i, \mathbf{p}_i) \rho(\mathbf{r}_i, \mathbf{p}_i) d\mathbf{r}_i^N d\mathbf{p}_i^N$	(2.2)
--	--	-------

In expressing \bar{O} in this way, one assumes that each point of the phase space can be visited with different occurrence probabilities. Thus, it is possible to define the time-independent probability density $\rho(\mathbf{r}_i, \mathbf{p}_i)$ characterizing each microstate. By definition, the probability density of the overall phase space is normalized to one.

Experimental measurements, on the other hand, provide the average of the same operator $O(\mathbf{r}_i, \mathbf{p}_i)$ over the measured time and over a macroscopic number of particles in the system. In other terms, the measured quantity is:

	$\bar{O} = \lim_{\tau \rightarrow \infty} \frac{1}{\tau - t_0} \int_{t_0}^{\tau} O\{\mathbf{r}_i(t), \mathbf{p}_i(t)\} dt$	(2.3)
--	--	-------

where $\{\mathbf{r}_i(t), \mathbf{p}_i(t)\}$ represent a trajectory $\{\Gamma(t)\}$ in phase space.

Needless to say, the trajectory is not explicitly known, and for a macroscopic sample (10^{23} particles), it would even be practically impossible to set suitable initial conditions, let alone solve the equations of motion.

The connection between the two points of view is established by the ergodic hypothesis,³⁸ stating that, for an isolated system and infinite sampling time, the trajectory will visit all microstates. If the ergodic hypothesis is fulfilled, time and ensemble averages coincide for any initial condition of the infinitely long trajectory. In the ensemble average, each microstate contributes to the integral by its probability density, $\rho(\mathbf{r}_i, \mathbf{p}_i)$.

2.1.1. The canonical ensemble

Under conditions of constant number of particles, N , system's volume, V , and temperature, T (canonical ensemble), the probability density is defined as:

	$\rho_{NVT}(\mathbf{r}_i, \mathbf{p}_i) = \frac{1}{h^{3N} N!} \frac{e^{-\beta H(\mathbf{r}_i, \mathbf{p}_i)}}{Q_{NVT}}$	(2.4)
--	---	-------

where h^{3N} is the quantized volume of a microstate, β is defined as $1/k_B T$, $H(\mathbf{r}_i, \mathbf{p}_i)$ is the Hamiltonian describing the kinetic and potential energy of the system, and $N!$ accounts for indistinguishable particles. A collection of configurations obeying to the probability density reported in Equation 2.4 is called Boltzmann distributed.

In this expression, Q_{NVT} is introduced simply as a normalization factor, since, by definition of probability distribution, the integral of $\rho_{NVT}(\mathbf{r}_i, \mathbf{p}_i)$ over phase space has to be equal to one. Hence:

	$Q_{NVT} = \frac{1}{h^{3N} N!} \int_{\mathbf{r}_i, \mathbf{p}_i} e^{-\beta H(\mathbf{r}_i, \mathbf{p}_i)} d\mathbf{r}_i^N d\mathbf{p}_i^N$	(2.5)
--	--	-------

In most cases, Q_{NVT} cannot be computed in practice using the Equation 2.5. However, Q_{NVT} , which is known as the partition function, plays a much larger role in statistical mechanics. In principle, once the partition function for an ensemble is known, all the macroscopic properties can be derived. First of all, within the canonical ensemble, the Helmholtz free energy, $F(N, V, T)$, is given by:

	$F(N, V, T) = U - TS = -k_B T \ln Q_{NVT}$	(2.6)
--	--	-------

where $U \equiv U(N, V, T)$ is the thermodynamic internal energy and $S \equiv S(N, V, T)$ is the system entropy. Moreover, using standard thermodynamic definitions and relations:

$S = -\left(\frac{\partial F}{\partial T}\right)_{V, N}$	$P = -\left(\frac{\partial F}{\partial V}\right)_{T, N}$	$\mu = -\left(\frac{\partial F}{\partial N}\right)_{T, V}$	(2.7)
--	--	--	-------

The link of macroscopic properties and partition function can be established also microscopically, since, for instance, the internal energy is given by:

	$U = \frac{1}{h^{3N} N! Q_N} \int H(\mathbf{r}_i, \mathbf{p}_i) e^{-\beta H(\mathbf{r}_i, \mathbf{p}_i)} d\mathbf{r}^N d\mathbf{p}^N = -\left(\frac{\partial \log Q_N}{\partial \beta}\right)_V$	(2.8)
--	--	-------

or

	$P = k_B T \left(\frac{\partial \log Q_N}{\partial V}\right)_{T, N}$	(2.9)
--	--	-------

In classical mechanics, the integration over momenta can be decoupled from the one over coordinates. Moreover, the integral over momentum of each particle can be computed independently and analytically, giving:

	$\frac{1}{h^3} \int d\mathbf{p} e^{-\beta \frac{p^2}{2m}} = \frac{1}{\Lambda^3}$	(2.10)
--	--	--------

where

	$\Lambda = \left(\frac{2\pi\beta\hbar^2}{m} \right)^{1/2}$	(2.11)
--	---	--------

is the de Broglie thermal wavelength of the particle.

In QM statistical mechanics, the \mathbf{r} and \mathbf{p} operators do not commute and $e^{-\beta\left[\frac{p^2}{2m}+U(\mathbf{r})\right]} \neq e^{-\beta\frac{p^2}{2m}} e^{U(\mathbf{r})}$. The exponential cannot be factorized and special methods (path integrals) are required.

In classical mechanics, the canonical partition function becomes:

	$Q_N = \frac{1}{N! \Lambda^{3N}} \int d\mathbf{r}^N e^{-\beta U(\mathbf{r}^N)} = \frac{Z_N}{N! \Lambda^{3N}}$	(2.12)
--	---	--------

where Z_N is the *configurational integral*:

	$Z_N = \int d\mathbf{r}^N e^{-\beta U(\mathbf{r}^N)}$	(2.13)
--	---	--------

In the ideal gas, $U(\mathbf{r}^N) = 0$ and $Z_N = V^N$, where V is the system volume. Hence, in the ideal gas case, the partition function is:

	$Q_N^{id} = \frac{1}{N!} \left(\frac{V}{\Lambda^3} \right)^N$	(2.14)
--	--	--------

and the free energy density, f^{id} , of the ideal gas is:

	$f^{id} = \frac{F^{id}}{V} = k_B T \rho (\log \rho \Lambda^3 - 1)$	(2.15)
--	--	--------

where ρ is the number density of particles, and the Stirling approximation $\log N! = N \log N - N$ has been used to express the factorial. The corresponding chemical potential, μ^{id} , is:

	$\mu^{id} = \left(\frac{\partial F^{id}}{\partial N} \right)_{V,T} = k_B T \log(\Lambda^3 \rho)$	(2.16)
--	---	--------

In the general case, it is easy to verify that the partition function can be re-written as:

	$Q_N = Q_N^{id} \frac{Z_N}{V^N}$	(2.17)
--	----------------------------------	--------

and the free energy is decomposed into ideal and excess contributions:

	$F = -k_B T \log Q_N = F^{id} + F^{ex}$	(2.18)
--	---	--------

where:

	$F^{ex} = -k_B T \log \left(\frac{Z_N}{V^N} \right)$	(2.19)
--	---	--------

2.1.2. The grand-canonical ensemble

All ensembles are equivalent in the thermodynamic limit. Nevertheless, it is sometimes useful to carry out computations in a specific ensemble different from the canonical one that we discussed in the previous paragraphs. In Chapter 4, use will be made of Monte Carlo in the grand-canonical ensemble, defined by the grand-potential, Ω , and by the partition function:

	$\Xi_{\mu VT} = \sum_{N=0}^{\infty} \frac{\exp[N\beta\mu]}{h^{3N} N!} \int \exp[-\beta H(\mathbf{r}^N)] d\mathbf{r}^N d\mathbf{p}^N = \sum_{N=0}^{\infty} \frac{z^N}{N!} Z_N$	(2.20)
--	---	--------

where z is the activity:

	$z = \frac{\exp[\beta\mu]}{\Lambda}$	(2.21)
--	--------------------------------------	--------

The grand-potential is given by:

	$\Omega(\mu, V, T) = -k_B T \log \Xi$	(2.22)
--	---------------------------------------	--------

with the further thermodynamic relations:

$S = -\left(\frac{\partial \Omega}{\partial T}\right)_{\mu, V}$	$P = -\left(\frac{\partial \Omega}{\partial V}\right)_{\mu, T}$	$N = -\left(\frac{\partial \Omega}{\partial \mu}\right)_{T, V}$	(2.23)
---	---	---	--------

Moreover, since $(z^N Z_N / N!)$ is the probability for the system to have N particles, it is possible to compute:

	$\langle N \rangle = \frac{\partial \log \Xi}{\partial \log N} = -\left(\frac{\partial \Omega}{\partial \mu}\right)_{T, V}$	(2.24)
--	---	--------

and, since the number of particles, fluctuates:

	$\frac{\langle (\Delta N)^2 \rangle}{\langle N \rangle} = \frac{\langle N^2 \rangle - \langle N \rangle^2}{\langle N \rangle} = \frac{k_B T}{\langle N \rangle} \frac{\partial \langle N \rangle}{\partial \mu}$	(2.25)
--	--	--------

Moreover, by definition:

	$N \left(\frac{\partial \mu}{\partial N} \right) = \frac{1}{\rho \chi_T}$	(2.26)
--	--	--------

where χ_T is the compressibility. Hence:

	$\frac{\langle (\Delta N)^2 \rangle}{\langle N \rangle} = \rho k_B T \chi_T$	(2.27)
--	--	--------

A relation of this type, giving a response function, χ_T , in terms of the fluctuation of an equilibrium quantity, can be seen as a special case of a “Green Kubo” relation or as the application of the fluctuation dissipation theorem.

In-depth discussions on the application of Monte Carlo in the grand-canonical ensemble are presented in Section 2.1.4.

2.1.3. The ideal free energy term

In our computations, we will deal with the absolute free energy of systems in the vapor, liquid, and solid phases, as well as with differences in free energy among samples in these three aggregation states. The computation of all these quantities involves a few subtleties that are discussed in this section.

Since, in most cases, we will work in the canonical ensemble, we start our discussion from the canonical partition function of a fluid system made of N simple particles:

	$Q_N = \frac{1}{N! h^{3N}} \prod_i \left(\int d\mathbf{p} \exp \left(-\frac{p^2}{2m_i} \right) \right) \int d\mathbf{r}^N \exp[-\beta U(\{\mathbf{r}^N\})]$	(2.28)
--	---	--------

From the N vector coordinates, \mathbf{r}_i , and momenta, \mathbf{p}_i , let us isolate the center of mass position and momentum:

	$\mathbf{R}_{CM} = \frac{1}{M} \sum_{i=1}^N \mathbf{r}_i$	$\mathbf{P}_{CM} = \sum_{i=1}^N \mathbf{p}_i$	(2.29)
--	---	---	--------

where $M = \sum_{i=1}^N m_i$.

Only $3(N - 1)$ independent coordinates and $3(N - 1)$ momenta remain that we can select to be Jacobi coordinates and momenta.³⁹ This choice is conceptually important, because we need to transform from a set of independent coordinates to a new set of equally independent coordinates, accounting from any Jacobian

factor that might arise in the transformation of integrals. In our computations, the separation of the center of mass motion will be carried out on a simple harmonic system, and here we do not need to go into details of the transformation of coordinates and momenta other than the center of mass ones.

Here, we only need to know that such a separation is possible and leaves behind $3(N - 1)$ *relative* coordinates and the corresponding $3(N - 1)$ momenta, that we indicate as $\{\mathbf{r}^{N-1}\}$ and $\{\mathbf{p}^{N-1}\}$.

Here, the partition function becomes:

	$Q_N = \frac{V}{Nh^3} \left(\int d\mathbf{P}_{CM} \exp\left(-\frac{P_{CM}^2}{2M}\right) \right) \times \frac{1}{(N-1)! h^{3N-3}} \prod_{i=1}^{N-1} \left(\int d\mathbf{p} \exp\left(-\frac{p^2}{2m_i}\right) \right) \int d\mathbf{r}^{N-1} \exp[-\beta U(\{\mathbf{r}^{N-1}\})]$	(2.30)
--	---	--------

where the volume factor results from the integration of \mathbf{R}_{CM} , taking into account that U does not depend on \mathbf{R}_{CM} .

The first line can be computed exactly, and the corresponding contribution to the free energy per particle is:

	$f_{CM} = k_B T \log \rho \Lambda_{CM}^3$	(2.31)
--	---	--------

where

	$\Lambda_{CM} = \left(\frac{2\pi\beta\hbar^2}{M} \right)^{1/2}$	(2.32)
--	--	--------

and $\rho = N/V$. The $(N - 1)!$ in the second line of Equation 2.30 will be computed by the Stirling approximation:

	$\log X! = X \log X - X$	(2.33)
--	--------------------------	--------

Hence, f_{CM} in Equation 2.31 represents the free energy due to the CM motion, while the second line in the partition function (Eq. 2.30) represents the free energy of a fluid system whose center of mass is kept fixed.

In the case of a solid sample such as the harmonic solid we use as a reference system, the partition function is slightly different:

	$Q_N = \frac{1}{h^{3N}} \prod_i \left(\int d\mathbf{p} \exp\left(-\frac{p^2}{2m_i}\right) \right) \int d\mathbf{r}^N \exp[-\beta U(\{\mathbf{r}^N\})]$	(2.34)
--	---	--------

since in this case particles are distinguished by the harmonic bonds, and the $N!$ factor is missing from the denominator. Notice that we write about *solid* samples and not *crystal* samples, since our systems, obtained in most cases by quench from the liquid, are invariably amorphous.

As a result of omitting $N!$ the partition function becomes:

	$Q_N = \frac{V}{h^3} \left(\int d\mathbf{P}_{CM} \exp\left(-\frac{P_{CM}^2}{2M}\right) \right) \times \frac{1}{h^{3N-3}} \prod_{i=1}^{N-1} \left(\int d\mathbf{p} \exp\left(-\frac{p^2}{2m_i}\right) \right) \int d\mathbf{r}^{N-1} \exp[-\beta U(\{\mathbf{r}^{N-1}\})]$	(2.35)
--	---	--------

The free energy of the center of mass motion is:

	$f_{CM} = k_B T \log \frac{\Lambda_{CM}^3}{V}$	(2.36)
--	--	--------

where Λ_{CM} is the same of fluid case.

This separation of the free energy of the center of mass motion from that of relative coordinates is very relevant for our story since in our simulations the center of mass of solid and fluid samples is kept fixed.

Moreover, in our computations, we will obtain the free energy of each system by first computing the free energy of a reference harmonic Hamiltonian, supplementing it with the perturbative estimate of the an-harmonic contribution to the free energy. The first contribution is computed from the harmonic frequencies, obtained in turn from the diagonalization of the dynamical matrix of the system. At this stage, the separation of the center of mass and of the relative coordinates is already implicitly made, reflected in the fact that the first three frequencies vanish. Thus, the harmonic free energy computed from the non-vanishing frequencies accounts for the contribution of relative coordinates. The perturbation step will concern only the difference in free energy due to the an-harmonic motion of relative coordinates. This subdivision of the computation in two steps is the reason why we could skip the details of the explicit transformation to Jacobi coordinates.

2.1.4. Molecular systems

Up to this point, for the sake of simplicity, all equations have been written for a single-component system made of simple (i.e. isotropic) particles.

In our case, we will deal with a mixture of molecular fluids or solids. In what follows, we will consider all molecules belonging to the same species as indistinguishable, while atoms within each molecule are distinguishable, since their relative position is fixed by covalent bonds that in our model cannot break.

Assuming that the system is made of N atoms organized into n_1 molecule of species 1, consisting of v_1 atoms, ..., n_q molecule of species q , consisting of v_q atoms, the canonical partition function becomes:

	$Q_N = \frac{1}{n_1! \dots n_q! h^{3N}} \int d\mathbf{p}^N \exp[-\beta KE(\{\mathbf{p}^N\})] \int d\mathbf{r}^N \exp[-\beta U(\{\mathbf{r}^N\})]$	(2.37)
--	---	--------

where KE is the system kinetic energy. The kinetic part can be factorized into the product of q factors, each representing the contribution of a single molecule:

	$Q_N = \frac{1}{n_1! \dots n_q! h^{3N}} \prod_{i=1}^q \left[\prod_{\alpha \in i} \int d\mathbf{p} \exp[-\beta \mathbf{p}^2 / 2m_\alpha] \right] \int d\mathbf{r}^N \exp[-\beta U(\{\mathbf{r}^N\})]$	(2.38)
--	---	--------

Moreover, in our force field model the potential energy term is the sum of q strong intra-molecular $\{U_i, i = 1, \dots, q\}$ terms, and a weaker inter-molecular contribution $\Delta U(\{\mathbf{r}^N\})$. Hence, we can re-write the partition function as:

	$Q_N = \frac{1}{n_1! \dots n_q! h^{3N}} \prod_{i=1}^q \left\{ \int d\mathbf{p} \exp \left[-\beta \sum_{\alpha \in i} \mathbf{p}^2 / 2m_\alpha \right] \int d\mathbf{r}^{v_i} \exp[-\beta U_i(\{\mathbf{r}^{v_i}\})] \right\} \\ \times \frac{\int d\mathbf{r}^N \exp[-\beta U(\{\mathbf{r}^N\})]}{\prod_{i=1}^q \int d\mathbf{r}^{v_i} \exp[-\beta U_i(\{\mathbf{r}^{v_i}\})]}$	(2.39)
--	--	--------

Since in each molecule atoms are distinguishable, the factor:

	$Q_i = \frac{1}{h^{3v_i}} \left\{ \int d\mathbf{p} \exp \left[-\beta \sum_{\alpha \in i} \mathbf{p}^2 / 2m_\alpha \right] \int d\mathbf{r}^{v_i} \exp[-\beta U_i(\{\mathbf{r}^{v_i}\})] \right\}$	(2.40)
--	--	--------

is the partition function of molecule i , contributing a term $f_i = -k_B T \log Q_i$ to the system free Helmholtz energy.

The remaining factor:

	$Q_N^{ex} = \frac{1}{v_1! \dots v_q!} \frac{\int d\mathbf{r}^N \exp\{-\beta [\sum_{i=1}^q U_i(\{\mathbf{r}^N\}) + \Delta U(\{\mathbf{r}^N\})]\}}{\prod_{i=1}^q \int d\mathbf{r}^{v_i} \exp[-\beta U_i(\{\mathbf{r}^{v_i}\})]} = \langle e^{[-\beta \Delta U(\{\mathbf{r}^{v_i}\})]} \rangle_{U_i}$	(2.41)
--	--	--------

summarizes the effect of the intra-molecular interaction. In this equation, $\langle \dots \rangle_i$ indicates the average on the trajectory generated by the intra-molecular Hamiltonian.

The canonical partition function for the system now is:

	$Q_N = Q_N^{ex} \prod_i Q_i$	(2.42)
--	------------------------------	--------

Within each molecule we can isolate a center of mass position, $\mathbf{R}_i = \sum_{\alpha \in i} \mathbf{r}_i / \sum_{\alpha \in i} m_i$, and the corresponding momentum, $\mathbf{P}_i = \sum_{\alpha \in i} \mathbf{p}_i$, leaving behind $3(v_i - 1)$ relative coordinates and $3(v_i - 1)$ momenta. Hence, each intra-molecular factor can be split into an ideal contribution arising from the center of mass motion, and intra-molecular coordinates relative to the center of mass.

In a similar way, the inter-molecular factor can be somewhat simplified, although at the cost of a slight approximation. The way to do this is to carry over to this factor the same subdivision of coordinates into center of mass (for each molecule) and relative coordinates. Then, let us consider that the intra-molecular potential energy terms do not depend on the center of mass coordinates (which are exact), and assume that the inter-molecular term depends only on the inter-molecular coordinates, identified by the molecular center of mass coordinates (which is approximated).

In this way, the integrals in Q_N^{ex} can be largely factorized, leaving:

	$Q_N^{ex} = \frac{1}{v_1! \dots v_q!} \int d\mathbf{R}_{CM}^N \exp[-\beta \Delta U(\{\mathbf{R}_{CM}^q\})]$	(2.43)
--	---	--------

Assuming that molecules are all the same, each contributing $f_{intra} = f_{vib} + f_{rot} + f_{id}$ to the system free energy, where f_{vib} , f_{rot} , and f_{id} are the vibrational, rotational and translational, and ideal intra-molecular free energies, respectively, the full partition function can be written as:

	$Q_N = \frac{1}{v!} \exp[-v\beta f_{intra}] \exp[-\beta f_{ex}]$	(2.44)
--	--	--------

The grand canonical partition function becomes:

	$\Xi_{\mu VT} = \sum_v \frac{1}{v!} \exp[-v\beta(f_{intra})] \exp[v\beta\mu] Q_v^{ex}$	(2.45)
--	--	--------

Equation 2.45 can be re-written as:

	$\Xi_{\mu VT} = \sum_v \frac{1}{v!} \exp[v\beta\Delta\mu] \int d\mathbf{R}_{CM}^v \exp[-\beta\Delta U]$	(2.46)
--	---	--------

where ΔU is the inter-molecular potential energy and $\Delta\mu = \mu - f_{vib} - f_{rot} - f_{id}$ is the μ arising from the inter-molecular interactions.

This relation is used to set up the grand canonical (GC) MC for the samples in the vapor phase (see Sec. 4.4.3) providing the chemical potential of molecules in the vapor. The aim of the MC computation is to estimate f_{ex} due to inter-molecular interactions, while f_{intra} is estimated from a harmonic computation for the single molecule whose center of mass is fixed at the origin.

Dealing with the configuration space, the potential energy of the system at each point can be defined as a single-valued function of coordinates, $U(\{\mathbf{r}^N\})$. Accounting for real systems including a large number of interacting particles, exhaustively sampling the entire configuration space would be unaffordable. Thus, only the most relevant configurations are usually probed, supposing that they have high probability, $\rho_{NVT}(\mathbf{r}_i)$, and contribute most to the ensemble average, $\langle O \rangle$. The most representative configurations are the ones of lowest energy. The two most popular methods used to generate Boltzmann distributed ensembles are the Monte Carlo algorithm and molecular dynamics simulation technique that are described in the next paragraphs. Note that if regions of high probability in configuration space are separated by significant energy barriers, it is unlikely that all the relevant configurations will be sampled by Monte Carlo or molecular dynamics techniques. Approaches to overcome this problem, known as quasi-non-ergodicity, rely on the so called enhanced sampling methods that are discussed later in this chapter.

From the knowledge of the potential energy, statistical mechanics provides a description of the system in terms of structure, dynamics, and time evolution at the conditions of interest. In addition, equilibrium and non-equilibrium properties can be defined. From here, my discussion will be focused on the potential energy.

2.2. Potential energy surface (PES)

Consider a diatomic molecule AB, whose structure is composed by two particles (atoms) connected by one spring (the chemical bond). This simplistic picture is the basis of the (classical) molecular mechanics (MM): stretching or compressing the springs, the molecular geometry is distorted leading the potential energy (depending on position) of the atomistic model to increase.

Quantum theory provides a better description of real molecules. Because of the uncertainty principle, quantum particles such as atoms cannot be localized at a single geometrical point, but vibrate incessantly about their equilibrium position. More precisely, they do not correspond to a unique position and momentum, but are described by a wave function. As a consequence, they always possess kinetic and potential energies also at $T \rightarrow 0$. Thus, in the quantum limit, a molecule is never stationary with zero kinetic energy and it always has zero point energy (ZPE).

Atomic interactions are described on the assumption that the potential energy of a system of N atoms is a single valued function of the $3N$ coordinates, $U \equiv U(\{\mathbf{r}^N\})$, defining its potential energy surface (PES).⁴⁰ More precisely, one usually refers to potential energy hypersurfaces, because of the high dimensionality of

the PES as a function of atomic coordinates. In the next sections, the concept of PES is introduced. Mathematical functions (i.e. force fields) helping the description of the behavior of the potential energy as a function of the system's geometric parameters are also described.

2.2.1. The Born-Oppenheimer approximation

The potential energy surface is defined as the system's potential energy as a function of the atomic coordinates. This simple and comprehensive description is possible thanks to the Born-Oppenheimer (BO) approximation.⁴¹ Born and Oppenheimer showed in 1927 that the Schrödinger equation for a molecule can be separated into an electronic and nuclear equation.

The starting point is the Hamiltonian for the combined electron and nuclei degrees of freedom:

	$\hat{H} = \hat{T}_{nuc} + \hat{T}_{ele} + U_{nuc-nuc} + U_{nuc-ele} + U_{ele-ele}$	(2.47)
--	---	--------

The Hamiltonian determines the system evolution through the time-dependent Schrödinger equation:

	$i\hbar \frac{\partial \Psi(\{\mathbf{r}_j\}; \{\mathbf{R}_I\} t)}{\partial t} = \hat{H} \Psi(\{\mathbf{r}_j\}; \{\mathbf{R}_I\} t)$	(2.48)
--	--	--------

or, equivalently, through its time-independent counterpart:

	$\hat{H}(\{\mathbf{r}_j\} \{\mathbf{R}_I\}) \Psi_\alpha(\{\mathbf{r}_j\}; \{\mathbf{R}_I\}) = E_\alpha \Psi_\alpha(\{\mathbf{r}_j\}; \{\mathbf{R}_I\}) \quad \alpha = 0, \dots, \rightarrow \infty$	(2.49)
--	---	--------

This same Hamiltonian, in turn, can be divided into nuclear and electronic terms:

	$\hat{H} = \hat{H}_{nuc} + \hat{H}_{ele}$	(2.50)
--	---	--------

The Hamiltonian of the electrons in the “external” Coulomb potential due to the nuclei is defined as follows:

	$\hat{H}_{ele} = \hat{T}_{ele} + U_{nuc-ele} + U_{ele-ele}$	(2.51)
--	---	--------

Because of the large ratio of nuclei and electrons mass, $M/m \geq 1800$, it is reasonable to assume that the electrons evolve faster than nuclei. In other words, the nuclei see the electrons as a cloud of negative charge which binds them in fixed relative positions thanks to the mutual attraction between positive and negative charges in the inter-nuclear region.

For every choice of coordinates of clamped nuclei, then, the focus is on the electronic problem, giving, in principle, an unlimited set of eigenvalues and eigenvectors, whose wave function is defined as:

	$\hat{H}_{ele}\psi_i(\{\mathbf{r}_j\} \{\mathbf{R}_I\}) = E_i(\{\mathbf{R}_I\})\psi_i(\{\mathbf{r}_j\} \{\mathbf{R}_I\}) \quad i = 0, \dots, \rightarrow \infty$	(2.52)
--	--	--------

The notation $(\{\mathbf{r}_j\}|\{\mathbf{R}_I\})$ means that the function depends on $\{\mathbf{r}_j\}$ given a set of $\{\mathbf{R}_I\}$ considered as additional parameters. The same assumption of fast evolution of electrons makes it plausible that they decay to their instantaneous ground state faster than the time scale of the nuclear motion. Hence, the only relevant electronic state is the ground state $i = 0$, and the electrons evolve *adiabatically* in the time-dependent field of the ions.

Replacing this information into the starting time-independent Schrödinger equation, and projecting $\Psi_\alpha(\{\mathbf{r}_j\}; \{\mathbf{R}_I\})$ onto $\psi_0(\{\mathbf{r}_j\}|\{\mathbf{R}_I\})$, one finds the Schrödinger-like equation satisfied by the wave function χ_α of the nuclei.⁴²⁻⁴³

	$[\hat{T} + E_0(\{\mathbf{R}_I\})]\chi_\alpha(\{\mathbf{R}_I\}) = E_\alpha\chi_\alpha(\{\mathbf{R}_I\})$	(2.53)
--	--	--------

where $\chi_\alpha(\{\mathbf{R}_I\}) = \int \Psi_\alpha(\{\mathbf{r}_j\}; \{\mathbf{R}_I\}) \psi_0^*(\{\mathbf{r}_j\}|\{\mathbf{R}_I\}) d\mathbf{r}_j$.

This equation describes the motion of nuclei on the *potential energy surface* $E_0(\{\mathbf{R}_I\})$. The decoupling of the electron and nuclear motion, and the identification of the potential energy surface $E_0(\{\mathbf{R}_I\})$ for the motion of nuclei is the fundamental result of Born-Oppenheimer. Under suitable conditions of low energy and temperature, the parameters $\{\mathbf{R}_I\}$ can be seen as the coordinates of classical particles. Under these conditions, we will write $U(\{\mathbf{R}_I\}) = E_0(\{\mathbf{R}_I\})$.

2.2.2. Stationary points of the PES

For all systems of interest in biophysics, the PES is a hyper-surface of really many dimensions, difficult to imagine and to visualize. A special role in describing its properties is played by the so-called *stationary points*, defined as the points such that the $3N$ -dimensional gradient $\nabla_{\mathbf{R}_I} U(\{\mathbf{R}_I\})$ vanishes.

To classify these stationary points, we will resort to the second derivative that can be organized into a matrix, which is known as the Hessian matrix.⁴⁴ In this definition, α, β label the Cartesian coordinates.

	$H_{I,J}^{\alpha,\beta} = \frac{\partial^2 U(\{\mathbf{R}_I\})}{\partial R_I^\alpha \partial R_J^\beta}$	(2.54)
--	--	--------

For all sufficiently regular $U(\{\mathbf{R}_I\})$, the matrix $H_{I,J}^{\alpha,\beta}$ is real and symmetric, hence it can be diagonalized giving $3N$ orthogonal eigenvectors, that, by convention, are normalized.

Supposing that the diagonalization of $H_{I,J}^{\alpha,\beta}$ has been carried out, we transformed to a basis of eigenvectors.

The Hessian define the first two terms in the multi-dimensional Taylor expansion of $U(\{\mathbf{R}_I\})$ around a stationary configuration $\{\bar{\mathbf{R}}_I\}$.

	$U(\{\mathbf{R}_I\}) = U(\{\bar{\mathbf{R}}_I\}) + \frac{1}{2} \sum_{J,K,\alpha,\beta} (R_J^\alpha - \bar{R}_J^\alpha) \frac{\partial^2 U(\{\bar{\mathbf{R}}_I\})}{\partial R_J^\alpha \partial R_K^\beta} (R_K^\beta - \bar{R}_K^\beta) + \dots$	(2.55)
--	---	--------

or, in the diagonal basis,

	$U(\{\mathbf{R}_I\}) = U(\{\bar{\mathbf{R}}_I\}) + \frac{1}{2} \sum_{J,\alpha} \frac{\partial^2 U(\{\bar{\mathbf{R}}_I\})}{\partial R_J^\alpha \partial R_J^\alpha} (R_J^\alpha - \bar{R}_J^\alpha)^2 + \dots$	(2.56)
--	--	--------

It is apparent that around all stationary points the behavior of $U(\{\mathbf{R}_I\})$ is determined by the sign and size of all eigenvalues of the Hessian matrix. When all eigenvalues are positive, $U(\{\mathbf{R}_I\})$ will grow in every direction. Hence, the configuration $\{\bar{\mathbf{R}}_I\}$ is a local minimum. The lowest among all local minima corresponds to the *ground state* configuration of the system.

When all eigenvalues are negative, we will have a maximum (not really relevant in this context).

The case of one negative eigenvalue and $(3N - 1)$ positive ones identifies a saddle point connecting the configuration basins around two distinct minima. The crossing between minima is particularly relevant in the transition state theory of reaction rates.

From the Hessian matrix, the vibrational frequencies characterizing the energy minimized molecular geometry can be computed. To this aim, the simplest vibrations of the molecule (i.e. normal-mode frequencies) need to be identified. In normal-mode vibrations, all the atoms move in phase with the same frequency. All vibrations of the molecule are the result of the combination of the normal modes. Consider a diatomic molecule, its normal-mode frequency, $\tilde{\nu}$, in cm^{-1} is defined as:

	$\tilde{\nu} = \frac{1}{2\pi c} \left(\frac{k}{\mu} \right)^{1/2}$	(2.57)
--	---	--------

where k is the force constant for the vibration (erg/cm^2), μ is the reduced mass of the molecule (g), and c , the velocity of light, is included to define the frequency, $\tilde{\nu}$, in cm^{-1} .

As reported in Equation 2.57, the frequency of a vibrational mode is related to the mode's force constant. Thus, the normal mode frequencies of a molecule (i.e. the directions and frequencies of the atomic motions), can be calculated from the force constant matrix. Indeed, from the diagonalization of the Hessian matrix, both the directions (eigenvectors) and the force constants (eigenvalues) can be obtained.

2.2.3. The force field

The potential energy surface is the crucial input to any atomistic or coarse grained simulation of condensed matter. Since every molecular dynamics or Monte Carlo simulation requires millions and even billions of energy evaluation, the PES has to be given by a mathematical expression that would be inexpensive to compute. Moreover, when MD is the method of choice, the PES expression has to be easy to differentiate to compute forces, and both energy and forces should be continuous.

Applications in organic chemistry and in biophysics exploit common features in the binding properties of these systems. Most biological systems are made of organic molecules, identified by a backbone of covalent bonds, interacting among themselves by relatively strong Coulomb interactions, and by weaker but pervasive dispersion interactions.

It was verified that the properties of covalent bonds are relatively transferrable from one molecular system to another, and relatively independent from the time-evolving geometry. Because of this observation, organic and biological systems can be seen as assemblies of particles (atoms, in most cases) and bonds, representing a network of springs with stretching, bending, and torsion energy contributions. Particles, moreover, may carry a Coulomb charge and interact with other particles by short range pair potentials representing dispersion forces. Short range repulsion due to Pauli's exclusion principle is also a general feature of real systems, and is modeled in all atomistic and coarse-grained force fields.

These qualitative considerations outline the *force field model* representing by far the most successful and most extensively used model to simulate biomolecules. In this approximation, the PES of the molecular system as a function of the nuclear coordinates is expressed as:⁴⁵

	$U_{FF} = \left(\sum_{1-2} U_{stretch} + \sum_{1-3} U_{bend} + \sum_{1-4} U_{tors} \right)_{bonded} + \sum (U_{elec} + U_{vdW})_{non-bonded}$	(2.58)
--	--	--------

In the following paragraph, each term included in the force field is briefly described.

Information on bonded interactions is provided by vibrational spectroscopy (infrared and Raman), since the stretching, bending, and torsion energy terms are fairly well distinct in energy, and the corresponding vibrational modes occupy different frequency bands in the vibrational spectrum.

The highest energy contribution is given by stretching, whose vibrational modes can reach up to 3600 cm⁻¹ in the case of C-H, N-H, and O-H bonds. Each stretching term, in particular, might be approximated fairly well by a Morse potential.⁴⁶

	$U_{stretch} = D_e [1 - e^{-\alpha(r-\bar{r})}]^2$	(2.59)
--	--	--------

In this expression, D_e is the well depth, α a parameter controlling the width of the potential, r is the inter-nuclear distance, and \bar{r} the equilibrium bond distance. For the most common covalent bond types, all these parameters can be computed by DFT probing large deviations of the inter-atomic distance away from the equilibrium one.⁴⁷ The Morse potential, however, is not commonly used because of the exponential function leading to computational inefficiency, the presence of three parameters to be defined, and the behavior of the function for $r \gg \bar{r}$. Since both compression and stretching of covalent bonds require a sizeable energy, at normal conditions the atomic motion occurs while keeping nearly constant the covalent bonding distance. Such a distance, moreover, is further enforced by quantum effects, freezing all modes whose energy exceeds the thermal energy. Then, for small deviations from the equilibrium bond distance, the stretching term can be written as a Taylor expansion in $(r - \bar{r})$. In its simplest form, the stretching term is usually approximated by the quadratic term of the Taylor expansion.

	$U_{stretch} = \frac{1}{2} k_{str} (r - \bar{r})^2$	(2.60)
--	---	--------

As a consequence of this choice, the validity of the model is restricted to near standard conditions (i.e. room temperature and atmospheric pressure), and the model could not be used to investigate phenomena such as bond breaking or formation.

Similarly to the stretching term, the angle bending contribution to the potential energy of the system is defined as:

	$U_{bend} = \frac{1}{2} k_{\theta} (\theta - \bar{\theta})^2$	(2.61)
--	---	--------

where $\bar{\theta}$ is the equilibrium bending angle.

The energy of bending bonds is also rather high compared to thermal energies. Thus, many models assume fixed bond angles.

Torsional angles are characterized by low internal rotation barriers, leading to the large variations of dihedral angles. The torsion energy as a function of the torsional angle is periodic through a 360° rotation. U_{tor} can be expressed with different functional forms depending on the atoms taken into account. Usually, the model relies on a short Fourier sum:

	$U_{tors} = \sum_{dihedrals} \sum_n \frac{V_n}{2} [1 + \cos(n\phi - \gamma)]$	(2.62)
--	---	--------

where V_n is the torsional rotation barrier, ϕ the dihedral angle, and γ the phase angles.

The number of terms to be included in the Fourier sum depends on the complexity of the torsional potential and the desired accuracy. In practical cases, n is taken up to 4.

In addition to the bonded terms, the force fields include the contributions of the non-bonded interactions to the system potential energy. Experimental information on these terms is made available by infrared spectroscopy for the Coulomb part and by thermodynamic functions for the dispersion terms.

The electrostatic interactions arise from differences in the charge distribution in a molecule. Assuming that point charges are placed at each atomic site and that for neutral molecules the sum is zero, the electrostatic term is usually modelled by a Coulomb potential:

	$U_{elec} = \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{i,j}}$	(2.63)
--	---	--------

where $q_{i,j}$ is the atomic charge, $r_{i,j}$ is the inter-nuclear distance between atoms i, j , and ϵ_0 is the vacuum permittivity.

In real systems, the charge on each atom depends on the local bonding environment, which is time-dependent. Moreover, atoms in condensed matter display polarization effect that change over time with changing system configuration. These effects are not explicitly accounted for by the so-called *rigid-ion* force fields, such as those used in the present thesis.

Finally, the Lennard-Jones (LJ) potential⁴⁸ is included in the force field to describe the van der Waals interactions accounting for all the non-electrostatic forces. The LJ potential accounts for both the short range repulsion and the medium range attractive dispersion terms, which depend on the inter-nuclear distance by $r_{i,j}^{-12}$ and $r_{i,j}^{-6}$, respectively. In the functional form reported in Equation 2.64, $\epsilon_{i,j}$ refers to the van der Waals well depth and $\sigma_{i,j}$ to the distance at which the potential is zero. The relationship between $\sigma_{i,j}$ and the minimum energy inter-nuclear distance, $r_{i,j}^0$ is also reported.

	$U_{vdW} = 4\epsilon_{i,j} \left[\left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^6 \right]$ $r_{i,j}^0 = 2^{1/6} \sigma_{i,j}$	(2.64)
--	--	--------

In most cases, the full set of $\epsilon_{i,j}$ and $\sigma_{i,j}$ is obtained from a more restricted set of homo-nuclear parameters $\epsilon_{i,i}$ and $\sigma_{i,i}$ through empirical equations, like the Berthelot's rule.⁴⁹

	$\epsilon_{i,j} = [\epsilon_{ii}\epsilon_{jj}]^{1/2}$ $\sigma_{i,j} = \frac{1}{2}[\sigma_{ii} + \sigma_{jj}]$	(2.65)
--	---	--------

Various force fields differ in their functional form and in the way their parameters were derived.⁵⁰ Bond lengths and angles parameters are often derived from high-level *ab initio* or density functional (DFT) calculations, or by crystal structures obtained by IR spectroscopy or X-ray crystallography. Torsional parameters can be adjusted to fit the profiles obtained from calculations or experiments. Dihedrals are usually fitted taking into consideration also the non-bonded interactions affecting the torsional barriers. A force field can be parametrized referring to a combination of quantum chemical calculations and experiments. Various force fields differ in the definition of the non-bonded parameters. For example, in AMBER⁵¹ and CHARMM,⁵² the charges are fitted to reproduce the electrostatic potential obtained from quantum mechanical calculations, whereas in OPLS⁵³ and GROMOS,⁵⁴ the non-bonded parameters are fitted to reproduce the thermodynamic properties.

In the last decades, a large effort was spent in the development of force fields suitable for modelling biological systems resulting in a good compromise between accuracy and computational efficiency.⁵⁵

However, the force fields are empirical and somewhat approximate. Some of the key approximations characterizing the force fields concern the use of fixed atomic charges preventing a proper description of charge polarizability. The need of a fixed system's topology prevents the applications of FF-based simulations to study chemical reactions, since chemical bonds cannot be broken and formed. In particular, it requires the assignment of specific protonation states to the molecular systems.

Among the available force fields, the Amber (Assisted Model Building with Energy Refinement) force field is one of the most applied for simulating proteins and nucleic acids.

2.2.3.1. The Amber force field

In 1984, Weiner et al.⁵⁶ developed a united-atom force field for simulating proteins and nucleic acids, incorporated into the Amber software package. In this model, equilibrium bond and angle parameters were obtained from crystal structures, whereas the dihedrals were adapted to match torsional barriers evaluated by means of both experimental data and quantum mechanical calculations. Charges were derived at the Hartree-Fock STO-3G level of theory. The van der Waals parameters were adapted from Hagler et al.⁵⁷ In 1986, the all atom version of the Amber force field was proposed.⁵⁸

Advances in computational resources made possible a more accurate parametrizations of the dihedrals and the partial charges, resulting in improved versions of the Amber force field.^{51, 59}

In 1995, Cornell et al.⁶⁰ introduced the set of parameters for all-atom simulations suitable for protein simulations in condensed phase, largely inspired by the OPLS potential.⁵³ In this force field, dubbed ff94 in AMBER, a new set of charges derived at the Hartree-Fock 6-31G* level of theory was introduced. A new set of van der Waals parameters was also introduced. Although these improved parameters resulted in a better description of long-range effects, an accurate treatment of electrostatic remained to be achieved. Indeed, the fixed point charges centered on the system atoms used by classical force fields may not be able to accurately describe the variability of the electrostatic properties to the system environment. To this aim, polarizable force fields have been introduced, but seldom used because the determination of multipolar interactions for all atoms requires at each time step iterative procedure that makes the model expensive and sometimes unstable.⁶¹

The structure of proteins is often described by a set of so-called ϕ, ψ dihedrals that were fit in ff94 to optimize the corresponding QM relative energies for several conformations of glycine and alanine, resulting in a poor parameterization of the protein backbone dihedral terms. Following versions of the Amber force field provided better parameterization of protein ϕ, ψ dihedrals. In the ff99SB force field, the ϕ, ψ dihedrals were parametrized by fitting the energies of multiple conformations of glycine and alanine tetra-peptides. Optimized side-chain torsions parameters were provided in the following versions of the force field improving the reproduction of experimental geometries, such as in one of the most recent versions (i.e. ff16SB).⁶²

The General Amber Force Field (GAFF)⁶³ was developed and included in the AMBER Antechamber program to parametrize small molecules.

Other widely used force fields include CHARMM, OPLS, and GROMOS. The usefulness of these force fields is greatly enhanced by the wide availability of computer packages optimized for massively parallel computer architectures, which have been developed to exploit them.

2.2.4. Density Functional Theory (DFT)

Biomolecules, like every other condensed matter system, can be seen as an assembly of electrons and atomic nuclei. Because of the BO approximation, nuclei can be treated as classical particles, while the electrons need to be described as quantum particles moving in the Coulomb field of nuclei.

Methods to compute the ground and excited states of N electrons in an external field are known, ranging from approximate approaches such as Hartree-Fock to exact methods such as configuration interactions. The complexity and cost of these high level methods limit their application to systems including a small number of atoms and electrons. Nowadays, the extensive study of biomolecules based on first principle relies on different approaches.

Density functional theory (DFT)⁶⁴⁻⁶⁵ acquires most of its computational efficiency by restricting its scope to ground state properties, and relies on the statement that for any N -electron system in an external field, the electronic ground state is the minimum of a universal functional of the electron density. The functional contains terms like the electron kinetic energy, the interactions of electrons with the M atomic nuclei, the mean-field Coulomb energy of the electron density (Hartree energy), as well as intrinsically quantum mechanical terms, such as the exchange and correlation energy, E_{XC} , which is itself a functional of the density.

	$E_{GS}[\rho \{\mathbf{R}_I, I = 1, \dots, M\}]$ $= E_k + \int \rho(\mathbf{r}) \left[\sum_I \frac{Z_I e^2}{ \mathbf{r} - \mathbf{R}_I } \right] d\vec{r} + \frac{e^2}{2} \int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{ \mathbf{r} - \mathbf{r}' } d\mathbf{r}d\mathbf{r}' + E_{XC}[\rho]$	(2.66)
--	--	--------

In the Kohn-Sham formulation, an auxiliary system of single-electron orbitals is introduced to compute the kinetic energy. These orbitals $\{\Psi_i(\mathbf{r}), i = 1, \dots, N\}$ are assumed to be orthonormal and the electron density is expressed as:

	$\rho(\mathbf{r}) = \sum_{i=1}^N \Psi_i(\mathbf{r}) ^2$	(2.67)
--	--	--------

The corresponding kinetic energy is defined as:

	$E_k = -\frac{1}{2} \sum_{i=1}^N \Psi_i^*(\mathbf{r}) \nabla^2 \Psi_i(\mathbf{r})$	(2.68)
--	--	--------

where the sum runs over singly occupied orbitals.

Atomic units have been used throughout.

Once the approximate expression is provided for E_{XC} , the functional can be minimized with respect to the orbitals giving the ground state energy and electron density.

The so-called Kohn-Sham (KS) equations are the result of the application of the variational principle to the functional. When supplemented by suitable boundary conditions, KS equations represent a set of coupled differential equations for the orbitals. These equations are self-consistent, since the Coulomb self-energy of the electron density and the exchange correlation potential depend on the resulting density. In practice the solution is obtained by iteration. First, one assumes a starting electron density, the Hartree and μ_{xc} potentials are constructed, and a first set of orbitals is computed. The corresponding density will be different from the input one. A new iteration is then performed with a new input density, which is a combination of the old

input and the output densities. The iterative procedure is continued until the output density differs from the input one less than a preselected threshold:

	$\int \rho_n(\mathbf{r}) - \rho_{n-1}(\mathbf{r}) d\mathbf{r} < \delta$	(2.69)
--	---	--------

where $\rho_n(\mathbf{r})$ and $\rho_{n-1}(\mathbf{r})$ are the output and input densities, respectively, and n is the iteration counter.

Density functional flavors are defined by the exchange-correlation approximation. A popular recipe is the generalized gradient corrected approximation exemplified by PBE (Perdew-Burke-Ernzerhof).⁶⁶ A hybrid exchange correlation functional, such as B3LYP,⁶⁷ usually is a combination of the Hartree-Fock exact exchange functional, $E_X[\rho(\mathbf{r})]$, and a generalized gradient term.

As previously described, DFT is a computational method to determine energy and electron density given a set of coordinates for the atomic nuclei. The extension of the method to optimize the system geometry and to perform molecular dynamics requires an improvement of the computational efficiency and stability. The discussion exceeds the scope of this section, and we refer to the book of Marx & Hutter⁶⁴ for a detailed discussion. Here we only mention that efficiency relies of the observation that molecular dynamics and geometry optimization require the computation of the energy on a sequence of closely related atomic configurations. In this case, most of the energy evaluations can be performed by updating the previous one. This can be achieved faster than starting an energy computation from scratch. This observation underlies most modern approaches to electronic structure and total energy computations by DFT, which are used in this thesis for accurate and predictive investigations of molecular species. From the point of view of these applications, DFT is simply an approach for the computation of the system potential energy surface. The underlying molecular dynamics machinery is purely classical.

2.3. The harmonic and quasi-harmonic approximation for molecular vibrations

The harmonic approximation (HA)⁶⁸ is an analytical theory introduced in condensed matter physics to compute thermodynamic and dynamical properties of solids and molecules. It can be formulated in quantum mechanical and classical terms. The simplicity and the low computational cost make the HA attractive to perform free energy calculations of systems at low temperature. With increasing temperature, as the system starts exploring the phase space around its minimum, an-harmonic effects become important, and the HA becomes inadequate. In those cases, corrections taking into account the third and fourth order of the Taylor series are applied as small perturbations on the dominant harmonic term.

A non-perturbative option is the quasi-harmonic approximation (QHA),⁶⁹ which identifies the minimum of the system free energy with respect to macroscopic variables like the volume.

Considering a Bravais lattice with a single atom basis where particles oscillate about their equilibrium (mean) positions, $\bar{\mathbf{R}}$, it is possible to define the atom displacement, $u(\bar{\mathbf{R}})$ such that $\mathbf{R} = \bar{\mathbf{R}} + u(\bar{\mathbf{R}})$, where \mathbf{R} is the instantaneous position. Therefore, a pair of atoms contributes an amount of $\Phi(\mathbf{R}) \equiv \Phi(\bar{\mathbf{R}} + u)$ to the potential energy, $U(\mathbf{R})$, of the system.

	$U(\mathbf{R}) = U(u) = \frac{1}{2} \sum_{\bar{\mathbf{R}}\bar{\mathbf{R}}'} \Phi(\mathbf{R} - \mathbf{R}') = \frac{1}{2} \sum_{\bar{\mathbf{R}}\bar{\mathbf{R}}'} \Phi(\bar{\mathbf{R}} - \bar{\mathbf{R}}' + u(\bar{\mathbf{R}}) - u(\bar{\mathbf{R}}'))$	(2.70)
--	---	--------

Thus, because of the dependence of the potential energy on the displacement, $u(\bar{\mathbf{R}})$, the total energy of the system can be stated by the Hamiltonian, H :

	$H = \sum_{\bar{\mathbf{R}}} \frac{P^2}{2M} + U(u(\bar{\mathbf{R}}))$	(2.71)
--	---	--------

where u are the coordinates, P the conjugate momenta, and M the atomic masses.

The potential energy, $U(\mathbf{R})$, can usually be expanded about the equilibrium position, $\bar{\mathbf{R}}$, as a Taylor series:

	$U(\mathbf{R}) = U(\bar{\mathbf{R}}) + \frac{dU(\bar{\mathbf{R}})}{d\mathbf{R}} u(\bar{\mathbf{R}}) + \frac{1}{2!} u(\bar{\mathbf{R}}) \frac{d^2 U(\bar{\mathbf{R}})}{d\mathbf{R}^2} u(\bar{\mathbf{R}}) + o(u(\bar{\mathbf{R}}))^3$	(2.72)
--	--	--------

In the framework of the harmonic approximation, the reference point, $\bar{\mathbf{R}}$, needs to be a local or global minimum of the potential energy surface of the system. Thus, the HA is built on any configuration such that forces on all atoms vanish:

	$\frac{dU(\mathbf{R})}{d\mathbf{R}} = 0$	(2.73)
--	--	--------

At low temperature, it is acceptable to expect that atoms bound by appreciable interactions will not deviate substantially from their equilibrium positions defining small displacements, $u(\bar{\mathbf{R}})$. These assumptions allow the truncation of the Taylor series to the second-order (i.e. quadratic form), leading to the following expression of the system potential energy:

	$U(\mathbf{R}) = U(\bar{\mathbf{R}}) + U_{\text{harm}}$	(2.74)
--	---	--------

where $U(\bar{\mathbf{R}})$ is the equilibrium potential energy and U_{harm} the harmonic potential energy that can be expressed as:

	$U_{harm} = \frac{1}{2} \sum_{\alpha, \beta} \sum_{i, j} \{u_i^\alpha\} \left(\frac{\partial^2 U}{\partial \mathbf{R}_i^\alpha \partial \mathbf{R}_j^\beta} \right) \{u_j^\beta\}$	(2.75)
--	---	--------

where i, j label atoms and α, β being Cartesian components (x, y, and z).

The expression of the system potential energy as a quadratic function of the atom displacements allows the analytical solution of the Newton's equations of motion reducing the problem to an eigenvalue problem. The Hessian matrix is strictly related to the harmonic force constant matrix including the atomic mass contributions (i.e. dynamical matrix). The vibrational frequencies, ω_i , and the vibrational eigenvectors of the system can then be obtained by the diagonalization of the dynamical matrix, $\frac{1}{\sqrt{m_i}} \frac{\partial^2 U}{\partial \mathbf{R}_i^\alpha \partial \mathbf{R}_j^\beta} \frac{1}{\sqrt{m_j}}$, which has real eigenvalues since it is symmetric. If $\bar{\mathbf{R}}$ is a minimum, as assumed here, the eigenvalues are positive. Upon diagonalization, the vibrational density of state (vDOS) can be computed characterizing the dynamical state of the system.

If U is the potential energy of an atomic or molecular solid, these considerations define the so-called *Debye model*.⁶⁸

Taking into account systems of N atoms, each of the $3N$ solutions of the eigenvalue problem describes one independent vibrational mode. If the energy does not change by displacing the system as a whole (no external field) three frequencies will vanish and their eigenvectors describe homogeneous rigid translations in the three directions. The $3(N - 1)$ other eigenvalues and eigenvectors describe independent vibrational modes characterized by a collective motion with a single frequency and constant phase. The $3(N - 1)$ eigenvectors are orthogonal with each other, and the elongation along normal modes can be used as generalized coordinates to describe the behavior of the system. In the framework of the harmonic approximation, non-linear molecules of N atoms are described in terms of independent harmonic oscillators, whereas $(3N - 5)$ normal modes are taken into account for describing linear molecules missing the rotation about the molecular axis.

Cartesian atom displacements and normal mode coordinates (which are again Cartesian, but possibly rotated with respect to the original ones) can be equivalently used in the definition of the system's Hamiltonian being related by an orthogonal linear transformation.

Given the vibrational frequencies, analytical expression for the free energy can be defined.

	$F = -k_b T \ln Q_{NVT} \quad \text{where } F(N, V, T)$ $G = -k_b T \ln \Xi_{\mu VT} \quad \text{where } G(\mu, V, T)$	(2.76)
--	--	--------

The Helmholtz free energy, F , requires the construction of the canonical partition function at constant temperature, Q_{NVT} , whereas the Gibbs free energy, G , is based on the grand canonical partition function at constant volume and chemical potential μ , $\Xi_{\mu VT}$.

The Hamiltonian of the harmonic oscillator of mass m and frequency ω is defined as:

	$\hat{H} = \frac{\mathbf{p}^2}{2m} + \frac{1}{2}m\omega^2\mathbf{r}^2$	(2.77)
--	--	--------

where \mathbf{r} and \mathbf{p} are the coordinate and conjugate momentum, respectively.

To account for the effect of the mass, one performs the canonical transformations of \mathbf{r} and \mathbf{p} as:

	$P = \frac{\mathbf{p}}{\sqrt{m}}$ $R = \mathbf{r}\sqrt{m}$	(2.78)
--	--	--------

Thus, the Hamiltonian of the harmonic oscillator of unitary mass can be defined as a function of the operators P and R :

	$\hat{H} = \frac{P^2}{2} + \frac{1}{2}\omega^2 R^2$	(2.79)
--	---	--------

whose equation of motion is

	$R(t) = A \cos \omega t + B \sin \omega t$	(2.80)
--	--	--------

In the quantum mechanical case, the partition function, Z , is defined as

	$Z = \sum_{n=0}^{\infty} \exp[-\beta E_n] = \sum_{n=0}^{\infty} \exp\left[-\beta\left(n + \frac{1}{2}\hbar\omega\right)\right] = \frac{\exp[-\beta\hbar\omega/2]}{1 - \exp[-\beta\hbar\omega]}$	(2.81)
--	---	--------

where $\beta = 1/k_B T$, k_B is the Boltzmann constant, $\hbar = h/2\pi$ is the Planck constant, and $E_n = \left(n + \frac{1}{2}\right)\hbar\omega$ represent the energy levels of the Hamiltonian.

Then, the Helmholtz free energy is defined as:

	$F = -k_B T \ln Z = \frac{\hbar\omega}{2} + k_B T \ln(1 - \exp[-\beta\hbar\omega])$	(2.82)
--	---	--------

where $\hbar\omega/2$ is recognizable as the zero point energy.

The entropy is then defined as:

	$\frac{S}{k_B} = -\frac{1}{k_B} \frac{\partial F}{\partial T} = \frac{\beta \hbar \omega}{\exp[\beta \hbar \omega] - 1} - \ln\{1 - \exp[-\beta \hbar \omega]\}$	(2.83)
--	---	--------

The partition function and thermodynamic properties of the classical oscillator can be obtained from the quantum mechanical expressions in the limit $\hbar \rightarrow 0$.

The classical partition function is defined as:

	$Z = \frac{k_B T}{\hbar \omega}$	(2.84)
--	----------------------------------	--------

where \hbar is the Planck constant, which is included in the definition as a factor transforming ω into the energy $\hbar\omega$, having the same dimensions of $k_B T$ and making Z dimensionless.

The classical Helmholtz free energy is then defined as:

	$F = -k_B T \ln Z = k_B T \ln \left(\frac{\hbar \omega}{k_B T} \right)$	(2.85)
--	--	--------

The entropy for the classical oscillator is

	$\frac{S}{k_B} = \ln \left(\frac{k_B T}{\hbar \omega} \right) + 1$	(2.86)
--	---	--------

Notice that the classical F and S have no finite limit for $T \rightarrow 0$. For this reason, in classical statistical mechanics, entropy and free energy have no natural reference value, and one deals only with the free energy differences at $T \neq 0$.

Here, two approximate methods for computing free energies of solids are mentioned: the Einstein approximation and the Debye approximation. Both methods were originally introduced to explain the low-temperature behavior of the heat capacities of solids.⁶⁸

The Einstein approximation assumes that the atoms of a crystal do not interact, and that they vibrate with the same phase harmonically and independently around a fixed center of mass. Thus, a simpler analytical expression for the free energy can be defined, involving a single frequency.

With the Debye model, one assumes that all the vibrational modes of a system can be obtained by the diagonalization of the dynamical matrix. The Debye approximation was originally introduced to describe the

heat capacity of solids considering only the three acoustic modes, which are those assumed to be linear and with the same slope. As a result of the Debye approximation, a quadratic expression for the vibrational density of state is applied to evaluate the system free energy.

To summarize, in the harmonic approximation, an interacting system is seen as a set of $3N$ harmonic oscillators or normal modes that vibrate independently of each other about their mean position. At higher temperatures, the vibrations of atoms in the harmonic system lead to populate states associate with higher energy while the mean positions of atoms remain unchanged. Thus, in the harmonic crystal, no thermal expansion is included. However, vibrations of a real system are not purely harmonic and at higher temperatures the mean positions of atoms change over time leading to thermal expansion.

Therefore, approaches taking into account the an-harmonicity of the system, such as the quasi-harmonic approximation (QHA), are needed to improve the description of systems made by interacting particles. The quasi-harmonic model extends the harmonic approximation to higher temperatures by introducing the dependence of frequencies on the system volume. An assumption of QHA is that the oscillations of the atoms in the system have harmonic-like frequencies that slightly change with changing of temperature and pressure. Thus, the harmonic approximation is valid for each volume, and the system free energy can be determined by the standard harmonic expression including the volume-dependent vibrational frequencies (Eq. 2.85). Therefore, the QHA takes implicitly into account the an-harmonic behavior of real system through the volume dependence of the atomic vibrations, giving access to a wider range of the equilibrium thermal properties of the system. For example, minimizing the system free energy, the equilibrium volume at any temperature can be extracted and the thermal expansion coefficient can be defined as:

	$\alpha_V(T) = \frac{1}{V(T)} \left(\frac{\partial V(T)}{\partial T} \right)_{P=0}$	(2.87)
--	--	--------

2.4. Molecular dynamics (MD)

Molecular dynamic (MD) is a method suitable for the local explorations of the configuration space. It allows the evaluation of the time evolution of a system and the calculation of time-dependent quantities.

The ability of MD to compute equilibrium properties relies on the ergodic theorem (Sec. 2.1).

Trajectories are the result of the numerical, step-by-step, integration of the classical Newton's equations of motion for a system represented as a set of atomic coordinates $\{\mathbf{r}_i, i = 1, \dots, N\} \equiv \{\mathbf{r}^N\}$ and conjugate momenta $\{\mathbf{p}_i, i = 1, \dots, N\} \equiv \{\mathbf{p}^N\}$. Thus, the classical equations of motion can be written as follows, where \mathbf{f}_i denotes the force acting on each i -th atom usually derived as the negative of the first derivative of the potential energy.

	$\mathbf{f}_i = \frac{d\mathbf{p}}{dt} = -\frac{\partial U(\mathbf{r})}{\partial \mathbf{r}}$	(2.88)
--	---	--------

Computing analytically the classical trajectory for a system of N particles would require solving a set of $3N$ coupled second order differential equations, making this approach inaccessible.

To overcome this problem, several approaches based on the time discretization of the evolution of the system were introduced. The finite difference methods allow the integration of the equation of motions in stages separated in time by a time step, δt , whose value has to be properly chosen to limit discretization errors. Usually, the choice of the timestep to apply during a MD simulation depends on the highest frequency motions of the system under investigation. Accounting for biological systems where the C-H stretching is the oscillator with the highest frequency (10 fs, 1 fs = 10^{-15} s), the time step is usually fixed at 1 fs. A common approach to increase the time step without altering the accuracy of the simulation relies on the application of constraints to some internal coordinates of the system by methods such as SHAKE.⁷⁰

Several algorithms, namely integrators, were developed for solving the equations of motion. They work approximating the positions, velocities, and accelerations as a Taylor expansion. Among the others, the Verlet integration scheme uses the information from the previous time step at time t to compute the new positions at time $t + \delta t$.⁷¹

	$\mathbf{x}(t + \delta t) = \mathbf{x}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 + o(\delta t^3)$	(2.89)
	$\mathbf{x}(t - \delta t) = \mathbf{x}(t) - \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 - o(\delta t^3)$	

where $\mathbf{a}_i = F_i/m_i$.

From previous equations, the following relation is obtained.

	$\mathbf{x}(t + \delta t) = 2\mathbf{x}(t) - \mathbf{x}(t - \delta t) + \mathbf{a}(t)\delta t^2 + o(\delta t^4)$	(2.90)
--	--	--------

The velocity at time t is then defined as:

	$\mathbf{v}(t) = [\mathbf{x}(t + \delta t) - \mathbf{x}(t - \delta t)]/2\delta t$	(2.91)
--	---	--------

Thus, for each particle of the system, the position at $t + \delta t$ is determined by the current position, $\mathbf{x}(t)$, the previous position, $\mathbf{x}(t - \delta t)$, and the acceleration, $\mathbf{a}(t)$, which is computed from the forces.

A variant of the Verlet algorithm is the velocity Verlet scheme,⁷² which uses a Taylor expansion truncated beyond the quadratic term for the coordinates. It first predicts coordinates and velocities at time, $t + \delta t$.

	$\mathbf{x}(t + \delta t) = \mathbf{x}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2$	(2.92)
	$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \mathbf{a}(t)\delta t$	

After the computation of energy and forces, the velocities are corrected.

	$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2}[\mathbf{a}(t + \delta t) - \mathbf{a}(t)]\delta t$	(2.93)
--	--	--------

In the velocity Verlet method, the positions, velocities, and accelerations at time $t + \delta t$ are obtained from the same quantity at time t . This is strictly needed whenever the underlying Hamiltonian or Lagrangian model requires velocities to define the forces, as it is the case, for instance, in the isobaric-isoenthalpic ensemble. Both Verlet and velocity Verlet methods achieve long term conservation of energy and momentum, if the time step is sufficiently short to ensure the stability of the integration.

In MD simulations, most of the time required for simulating a large system is used to evaluate non-bonded interactions. The number of bonded interactions scales linearly with the system size. To approach linear scaling also with non-bonded interactions, a neighboring list is used to make linear the computation of the short-range repulsion and dispersion forces. Coulomb contributions are evaluated by the Ewald sum⁷³ for systems with up of a few thousand particles, and by its mesh-based versions⁷⁴ for larger systems.

Periodic boundary conditions are usually set up to reduce the effect of finite size and to approximate the infinite system by the simulated one.

The ensemble described by the Hamiltonian dynamics of MD is the micro-canonical (NVE) ensemble defined by fixing the number of particles, N , the volume, V , and the total energy of the system, $H(\mathbf{r}, \mathbf{p}) = K(\mathbf{p}) + U(\mathbf{r})$. The micro-canonical ensemble allows the exploration of the conformational space at constant total energy. However, if the results from simulations are compared with experimental data, the NVE ensemble finds little application, because the experiments are carried out at constant pressure and temperature.

Extending the integration of Newton's equations of motion to the isothermal-isobaric (NPT)⁷⁵ or the canonical (NVT)⁷⁶ ensembles is possible. In those cases, the system total energy is not a constant of motion and the resulting dynamics is defined as non-Hamiltonian.

To this aim, an extended bath of constant pressure/temperature, defined as a barostat/thermostat, respectively, is coupled to the simulated system. A thermostat is an algorithm changing the equations of motion in order to obtain the distribution of microstates compatible with the probability density of the canonical ensemble, $\rho_{NVT}(\mathbf{r}_i, \mathbf{p}_i)$. The employment of a thermostat requires the definition of the

instantaneous temperature, which is directly related to the particles' kinetic energy. Different thermostats are available, such as the Nosé-Hoover,⁷⁶⁻⁷⁷ the Berendsen,⁷⁸ the Langevin,³⁸ and the Bussi-Parrinello⁷⁹ thermostats. Stochastic dynamics can also be applied for this purpose.⁸⁰ Similar considerations can be done regarding barostats algorithms,⁴⁰ controlling the system pressure by scaling the volume.

2.5. Monte Carlo (MC) methods

Monte Carlo (MC) was first introduced in applied mathematics as a method to estimate integrals over multi-dimensional domains.⁸¹ The method consists of generating a number of random points, x_i , accordingly to a pre-assigned non-uniform probability distribution, $\rho(x_i)$. The application of MC to statistical mechanics was driven by the multidimensional integrals defining averages over phase space, $\langle \hat{O} \rangle$.

In the canonical ensemble, the expectation value of $\langle \hat{O} \rangle$ may be estimated through the importance sampling of \hat{O} over the distribution, $\rho(\mathbf{r})$, defined as follows.

	$\rho(\mathbf{r}) = \frac{e^{-\beta U}}{\int e^{-\beta U} d\mathbf{r}_i^N}$	(2.94)
--	---	--------

In most of the relevant cases, the integrand $e^{-\beta U}$ is vanishingly small nearly everywhere, and no direct placement of points over a multidimensional space according to the Boltzmann distribution can achieve a sufficient efficiency. In addition, the integral representing the normalization of the probability distribution is unknown.

Thus, a practical way to achieve importance sampling over a non-normalized distribution is provided by the Metropolis algorithm.⁸² The method relies on the generation of a sequence of points (configurations), whose distribution progressively approaches a pre-assigned probability density proportional to $e^{-\beta U}$. The current position is updated without memory of previous steps, defining a Markov chain. Each step in the MC evolution consists of an attempted move from the present state $\{\mathbf{r}\}$ to $\{\mathbf{r}'\}$ (close to $\{\mathbf{r}\}$), followed by an acceptance-rejection decision.

In the canonical ensemble, the new configuration $\{\mathbf{r}'\}$ is accepted with the probability, ρ , defined as follows.

	$\rho(\mathbf{r} \rightarrow \mathbf{r}') = \min\{1, \exp[-\beta(U(\mathbf{r}') - U(\mathbf{r}))]\}$	(2.95)
--	--	--------

Thus, the rules used to generate the sequence of points require only the ratio of probabilities for different states, making unnecessary the explicit knowledge of the normalization of the probability distribution.

For a detailed discussion on MC methods, I refer to the work of Hammersley.⁸¹

In comparison to MD, MC allows the exploration of the configuration space without requiring the computation of forces. Hence, discontinuous or even hard-core potentials can be used. In addition, significantly different volume regions of configuration space can be explored introducing ad-hoc MC moves. On the other hand, the trajectories generated by a MC simulation do not reproduce the real time dynamics of the system and no time correlation function can be computed. Moreover, each MC step, although faster to execute than a single MD step, moves less in phase space.

Like MD, MC can be extended to other ensembles.

In the computations presented in Chapter 4, use will be made of the grand-canonical (GC) formulation (Sec. 2.1.2 for details on the GC ensemble and Sec. 2.1.4 for details on the application to molecular systems). In this approach, the standard sampling of Monte Carlo is supplemented by attempts to add or remove one particle from the system. Since the probability distribution for the system of having N particles in it is defined as:

	$\rho(N) = \frac{1}{\Xi \Lambda^{3N}} \frac{e^{\beta N \mu} Z_N}{N!}$	(2.96)
--	---	--------

where μ is the chemical potential and Z_N the configurational partition function, the probability of adding one particle is:

	$\alpha(N \rightarrow N + 1) = \frac{V \Lambda^{-3}}{(N + 1)} \exp[\beta \mu] \exp \left[-\beta \left(U(\{\mathbf{r}^{(N+1)}\}) - U(\{\mathbf{r}^{(N)}\}) \right) \right]$	(2.97)
--	--	--------

while the probability of removing one particle is:

	$\alpha(N \rightarrow N - 1) = \frac{N}{V \Lambda^{-3}} \exp[-\beta \mu] \exp \left[-\beta \left(U(\{\mathbf{r}^{(N+1)}\}) - U(\{\mathbf{r}^{(N)}\}) \right) \right]$	(2.98)
--	---	--------

where the $1/V$ factor has been introduced to account for the uniform probability of placing (or removing) one molecule at any given position in the simulation cell.

The application of GC-MC to compute the chemical potential and thus free energies of molecular systems requires some elaborations already anticipated in Section 2.1.4. For efficiency reason, the GC-MC applications concern gases ($T > T_c$) and vapors ($T_{boil} < T < T_c$), excluding liquids and solids. Since in our computations GC-MC is used to simulate relatively dilute vapors made of fairly rounded molecules, the system will be seen as an assembly of weakly interacting complex particles of free energy, $f_{intra}(T)$ arising from intra-molecular forces, and computed according to the method detailed in Chapter 4. The probability of inserting or removing a whole molecule already equilibrated at temperature, T , now is defined as reported in Equations 2.99 and 2.100, respectively.

	$\alpha(N \rightarrow N + 1) = \frac{V\Lambda^{-3}}{(n + 1)} \exp[\beta(\Delta\mu)] \exp \left[-\beta \left(\Delta U(\{\mathbf{R}^{(n+1)}\}) - \Delta U(\{\mathbf{R}^{(n)}\}) \right) \right]$	(2.99)
	$\alpha(N \rightarrow N - 1) = \frac{n}{V\Lambda^{-3}} \exp[-\beta(\Delta\mu)] \exp \left[-\beta \left(\Delta U(\{\mathbf{R}^{(n+1)}\}) - \Delta U(\{\mathbf{R}^{(n)}\}) \right) \right]$	(2.100)

where n is the number of molecules, and $\Delta U(\{\mathbf{R}^n\})$ is the inter-molecular potential energy, dependent on the positions $\{\mathbf{R}^n\}$ of the molecular centers of mass. The factor $1/V$ is included to account for the homogeneous probability distribution of inserting or removing a molecule in the simulation volume. The chemical potential μ in these equations accounts for the inter-molecular interactions.

This formulation allows us to focus on the $\Delta\mu$ parameter accounting for inter-molecular interactions, which can be computed very accurately, and added to any other estimate of the free energy difference between the molecule in solution and in the reference state.

In practice, a MC is run for an assembly of n molecules, considering:

- Single atom moves;
- Whole molecule rotations;
- Rotations of molecular sub-units, whenever suitable sub-units (such as $-\text{CH}_3$ in isobutane) can be identified.

At every step, the insertion of a molecule is attempted with probability α , with $1/\alpha \sim 100 - 1000$. This move consists of cloning a molecule already in the system, and thus already equilibrated, placing it at a random position, accepting or rejecting with the probability reported in Equation 2.99.

In a similar way, again with probability α per step, a molecule is selected at random, and its removal from the system is attempted with the probability reported in Equation 2.100.

In our study, the goal of the simulation is to determine the value of μ that corresponds to the experimental density of the vapor. Then, the chemical potential μ is equated to the Helmholtz free energy, f_{inter} per molecule due to inter-molecular interactions. This is an approximation since $\mu = \partial F / \partial N$, but this term is already fairly small, and we will neglect the small error due to the approximation. Moreover, $\mu = G/N$, and at the low pressure of the simulations, $G \sim F$, which is a very accurate approximation.

2.6. Rare events

Many events of interest for bio-physics and bio-chemistry, such as large conformational changes or absorption/release of a ligand by a receptor, are usually referred to as rare events. Those phenomena are

characterized by states that, at equilibrium, are separated by energy barriers leading to long waiting time for the spontaneous event to occur. Thus, the rare event takes place in a fraction of the time required to investigate the phenomenon. The need to cover a long time to analyze a very short-lasting event makes this investigation a hard task. Algorithms to overcome this problem have been developed, but their application to large biological systems is still challenging.

In this context, the Transition State Theory (TST) is briefly introduced.

2.6.1. Transition state theory (TST)

A rare event requires the crossing of an energy barrier moving from the reactant to the product. Assuming that the product is depopulated, the Arrhenius equation⁸³ can be defined accounting for the temperature dependence of the transition rate, k .

	$k = A \exp \left[-\frac{E_a}{k_B T} \right]$	(2.101)
--	--	---------

More importantly, it introduces the idea of an activated state energy, E_a , that the reactants have to cross to reach the product state.

Transition state theory (TST) has been developed in 1935 by Eyring, Wigner, Evans, and Polanyi.²¹ The exponential dependence on temperature of the Arrhenius equation is incorporated into the TST of reaction rates. In TST, the reaction rate is estimated combining concepts of thermodynamics, collision theory, and statistical mechanics, and it is defined as:

	$k = \frac{k_B T}{h} \exp \left[-\frac{\Delta H^* - T \Delta S^*}{k_B T} \right]$	(2.102)
--	--	---------

where ΔH^* , ΔS^* are the enthalpy and entropy of the transition state relative to the local equilibrium in the reactant's basin, h is the Planck's constant, and k_B is the Boltzmann's constant.

In the framework of TST, reactants and products are macroscopic states, corresponding to basins in the system phase space. This condition is referred to as *near equilibrium*, in which the transition state is in equilibrium with the reactants, and the transition is complete only when the system equilibrates in the product basin. A relatively high energy barrier separates the reactants to the products, defining a sparsely populated region of the phase space (i.e. the transition state).

Accounting for systems made of many particles, the potential energy surface (PES) has to be taken into account. To identify the transition state, a transition path describing the progress of the system toward the product, needs to be identified. A steepest descent line on the PES running between the two basins is the first obvious choice for the transition path, labelled by a 1D variable, $s(\mathbf{r}^N)$, known as reaction coordinate. The

transition state is then identified as the point along the path characterized by the highest potential energy. Being defined by a unique value of the reaction coordinate, one deals with a $(6N - 1)$ -dimensional hyper-surface in phase space. In the original $6N$ -dimensional space, the transition state is identified as the minimum energy point of the $(6N - 1)$ -dimensional transition hyper-surface, corresponding to a saddle point in the $6N$ -dimensional space. The system geometry compatible to the transition state is the closest analogue that TST may define in terms of a transition configuration. By integrating $e^{-\beta H}$ over the transition state hypersurface, the Helmholtz free energy of the $(6N - 1)$ subspace can be determined. Once measured from the reactant basin, this free energy represents the $(\Delta H^* - T\Delta S^*)$ transition free energy barrier determining the transition rate.

Thus, TST defines the reaction rate as the equilibrium flux of the system across the $(6N - 1)$ -dimensional transition hyper-surface. However, in this definition the re-crossing of the same transition barrier until the equilibrium with the product basin is established, is not taken into account. As a consequence, the TST rate is an upper bound to the true underlying rate. To minimize and even prevent the re-crossing, the variational transition state theory (VTST) has been introduced.⁸⁴

2.7. Free energy calculations

Experimental results often represent, or depend on, free energy differences. At equilibrium, these may represent the change in the system free energy when: a ligand binds to or unbinds from a protein; a chemical reaction takes place; or external conditions such as temperature and pressure are varied. Free energy differences control also the kinetics of rare events in the system, since the activation free energy represent the increase of the system free energy from the starting condition to the transition state towards a different stationary state.

The normal formulation of statistical mechanics is also tuned to the computation and interpretation of free energy differences, although practical applications are still fraught of technical and computational difficulties.

The basic definition of the Helmholtz free energy in terms of the canonical partition function Q_{NVT} is:

	$F = -k_B T \ln Q_{NVT}$	(2.103)
--	--------------------------	---------

where k_B is the Boltzmann's constant.

On the other hand, as stated in the previous sections:

	$Q_{NVT} = \frac{1}{N! h^{3N}} \int_{\mathbf{r}_i, \mathbf{p}_i} e^{-\beta H(\mathbf{r}_i, \mathbf{p}_i)} d^{3N} \mathbf{r}_i d^{3N} \mathbf{p}_i$	(2.104)
--	--	---------

where h^{3N} is the quantized volume of a microstate, and β is defined as $1/k_B T$. The Hamiltonian, $H(\mathbf{r}_i, \mathbf{p}_i)$, gives the total energy of the system in a given configuration (i.e. a set of coordinates and momenta) accounting for both the kinetic and potential energy.

Equations 2.103 and 2.104 are easily adapted to the computation of free energy differences.

Let us assume that A and B differ by some detail of their Hamiltonians H_A and H_B . Since, in most cases, the difference concerns the potential energy part, we define:

	$\Delta H_{BA} = H_B - H_A = \Delta U_{BA} = U_B - U_A$	(2.105)
--	---	---------

Hence, the free energy difference ΔF_{BA} is written as:

	$\Delta F_{BA} = F_B - F_A = -k_B T \ln \frac{Q_B}{Q_A}$	(2.106)
--	--	---------

Provided A and B containing the same number of particles, the integral over momenta cancels out and ΔF_{BA} can be equivalently expressed as a function of the configurational integrals, Z_A and Z_B :

	$\Delta F_{BA} = F_B - F_A = -k_B T \ln \frac{Z_B}{Z_A}$	(2.107)
--	--	---------

In the context of this thesis, free energy differences are relevant in two major cases. In the first case, we will be interested in the free energy difference between two real systems, differing, as already stated, by their Hamiltonian, by their configuration, or because their thermodynamic parameters such as temperature or pressure are different. In the second state, particularly important for the developments presented in Chapter 4, the difference concerns a real (i.e. fully interacting) system and an auxiliary one, namely the *reference model*, somewhat simpler and amenable to analytical solution. The latter case represents the most suitable way to compute absolute free energies, provided the free energy of the reference model is known.

Both cases are dealt by computational approaches based on the logarithmic expression for F (Eq. 2.103) possibly combined with the explicit definition of partition function. In one case, one obtains an expression for F_{BA} directly in terms of an average over the distribution functions ρ_A or ρ_B of states A and B; in the second case, F_{BA} is obtained by integrating the derivative $dF(\lambda)/d\lambda$ with respect to the parameter λ morphing A into B or vice-versa.

2.7.1. Free energy perturbation theory

According to thermodynamic perturbation theory, or, in this case, free energy perturbation (FEP), the difference in free energy between states A and B is related to the average of the exponential of the energy difference between these states, $\Delta U_{BA} = U_B(\mathbf{r}) - U_A(\mathbf{r})$:⁸⁵

$ \begin{aligned} \Delta F_{BA} &= -k_B T \ln \frac{Z_B}{Z_A} = -k_B T \ln \frac{\int \exp[-\beta U_B(\mathbf{r})] d\mathbf{r}}{\int \exp[-\beta U_A(\mathbf{r})] d\mathbf{r}} \\ &= -k_B T \ln \frac{\int \exp[-\beta (U_B(\mathbf{r}) - U_A(\mathbf{r}))] \exp[-\beta U_A(\mathbf{r})] d\mathbf{r}}{\int \exp[-\beta U_A(\mathbf{r})] d\mathbf{r}} \\ &= -k_B T \ln \int \exp[-\beta (U_B(\mathbf{r}) - U_A(\mathbf{r}))] \rho_A(\mathbf{r}) d\mathbf{r} \\ &= -k_B T \ln \langle \exp[-\beta (U_B(\mathbf{r}) - U_A(\mathbf{r}))] \rangle \end{aligned} $	(2.108)
---	---------

where $\rho_A(\mathbf{r})$ is the configurational probability density function corresponding to state A. Of course, a similar relation can be derived based on the distribution function $\rho_B(\mathbf{r})$:

$\Delta F_{AB} = -k_B T \ln \langle \exp[-\beta \Delta U_{BA}] \rangle_B$	(2.109)
---	---------

Thus, the free energy difference ΔF_{AB} can be obtained from a single simulation of state A (or, equivalently, B). The ensemble average of Equation 2.109 is dependent on the exponential factor, $\exp[-\beta \Delta U_{BA}]$. Thus, only the configurations characterized by the lowest ΔU_{BA} values will contribute significantly to the ensemble average. However, those configurations might fall (as they often do) where $\rho_A(\mathbf{r})$, also dependent on an exponential function, is vanishingly small. Thus, it is apparent that the practical application of these relations require the distributions $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$ to overlap.

Therefore, approaches based on simulating both states A and B were proposed. In such methods, the free energy difference, ΔF_{BA} , is derived from the intersection point of the energy difference distributions of states A and B, $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$. To overlap the distributions, the transformation is split in windows and the free energy change is computed summing the free energy differences between subsequent windows. The free energy difference distributions are then estimated by constructing a histogram. Note that only accurate estimates of the energy difference distributions would allow the reconstruction of the free energy difference with sufficient accuracy.

Another approach to estimate the difference in free energy using computer simulations was proposed by Bennett.⁸⁶ This method is based on an iterative process aimed to finding the optimal intermediate state, R , corresponding to the state whose energy difference distribution, $\rho_R(\Delta U_{BA})$ is the one showing the highest

probability in the intersection region of $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$. As a consequence, the free energy difference may be estimated from a single simulation of the optimal intermediate state.

2.7.2. Thermodynamic integration

Thermodynamic integration (TI) was first introduced by Kirkwood.⁸⁷ The method requires the definition of an unphysical path connecting states A and B through an intermediate Hamiltonian depending on the perturbation parameter, λ , and of a sampling scheme based on several independent simulations at different λ values. Thus, the difference in free energy is computed as the integration over the whole λ range of the derivative of the free energy with respect to the perturbation parameter, λ .

	$\Delta F_{BA} = F_B - F_A = F(\lambda_B) - F(\lambda_A) = \int_{\lambda_A}^{\lambda_B} \frac{dF}{d\lambda} d\lambda$	(2.110)
--	---	---------

Since the excess free energy can be expressed as:

	$F^{exc} = -\frac{1}{\beta} \ln Z(\lambda)$	(2.111)
--	---	---------

One obtains:

	$\frac{dF^{exc}(\lambda)}{d\lambda} = -\frac{1}{\beta Z(\lambda)} \left(\frac{\partial Z(\lambda)}{\partial \lambda} \right)$	(2.112)
--	--	---------

where the configurational partition function $Z(\lambda)$ is defined as:

	$Z(\lambda) = \int d\mathbf{r}^{3N} e^{-\beta U(\lambda)}$	(2.113)
--	--	---------

As a consequence, one obtains:

	$\frac{\partial Z(\lambda)}{\partial \lambda} = -\beta \int d\mathbf{r}^{3N} \frac{\partial U(\lambda)}{\partial \lambda} e^{-\beta U(\lambda)}$	(2.114)
--	--	---------

In most applications, $\partial U(\lambda)/\partial \lambda = \partial H(\lambda)/\partial \lambda$.

Hence:

	$\frac{dF^{exc}(\lambda)}{d\lambda} = \frac{\int d\mathbf{r}^{3N} \frac{\partial U(\lambda)}{\partial \lambda} e^{-\beta U(\lambda)}}{\int d\mathbf{r}^{3N} e^{-\beta U(\lambda)}} = \langle \frac{\partial H(\lambda)}{\partial \lambda} \rangle_\lambda$	(2.115)
--	--	---------

As already stated, the computation of absolute free energies and of relative free energies has deep similarities. Differences, however, arise at the algorithmic level, due to reciprocal advantages and disadvantages of the two cases.

2.7.3. Relative free energy determination

As pointed out by Christ et al.⁸⁸ to accurately compute free energy differences by molecular simulations, the Hamiltonian describing the total energy of the system, the scheme to sample the configuration space, and the free energy estimator have to be chosen carefully. The Hamiltonian used to evaluate the energy and forces must be able to reproduce all the configurations with the correct relative probability in a reasonable simulation time. The computational efficiency strongly depends on the system degrees of freedom and on the functional form of the Hamiltonian. A reduction of the degrees of freedom can be obtained moving from a quantum-mechanical description of the system where the electronic degrees of freedom are modeled explicitly, to a classical one where the atom is treated as a single particle or to a coarse-grained model in which groups of atoms are merged into one particle. The reduction of the system size and the treatment of part of the system as a continuum would further reduce the system's degrees of freedom. In molecular simulations, the functional form of the classical Hamiltonian relies on the force field.

To compute the free energy difference between two states, A and B, however, the overlap region between the probability densities of states A and B have to be sufficiently sampled. This is shown by the following derivation.

Let us define $\rho_A(\Delta H|\Delta H_{BA})$ as the probability for the energy difference $\Delta H_{BA} = H_B - H_A$ to be equal to ΔH :

	$\rho_A(\Delta H \Delta H_{BA}) = \frac{\int d\mathbf{r} d\mathbf{p} e^{-\beta H_A} \delta(\Delta H - H_{BA})}{\int d\mathbf{r} d\mathbf{p} e^{-\beta H_A}}$	(2.116)
--	--	---------

The probability density $\rho_B(\Delta H|\Delta H_{BA})$ is defined in a similar way.

Then, by simple algebraic manipulations, it is possible to show that:

	$\rho_B(\Delta H \Delta H_{BA}) \exp[-\beta \Delta F_{BA}] = \rho_A(\Delta H \Delta H_{BA}) \exp[-\beta \Delta H]$	(2.117)
--	--	---------

where the free energy difference is defined as $\Delta F_{BA} = F_B - F_A$ and ΔH is the difference in energy.

Equation 2.117 expresses that at the point where the two probability densities, ρ_A and ρ_B , are the same, the free energy difference ΔF_{BA} is equal to the difference in energy ΔH . Therefore, sampling the overlap region of states A and B is crucial to obtain accurate (converged) free energy estimates.

As already mentioned, Monte Carlo⁸² and molecular dynamics are widely used simulation methods able to perform local explorations of the configuration space. However, because of the inadequate sampling of the

overlap region, an intermediate Hamiltonian might be defined as a function of a coupling parameter, λ , joining the state A ($\lambda=0$) to state B ($\lambda=1$).⁸⁹⁻⁹⁰ Being the free energy a state function, the dependence of the Hamiltonian on λ can vary without changing the difference. In the application presented in this thesis, a linear combination of Hamiltonians was defined. The choice of the combined Hamiltonian has a critical influence on the efficiency of the free energy estimate.

Once defined the combined Hamiltonian, the sampling scheme has to be set up in order to probe all important configurations (i.e. the configuration at lower energy) over the whole λ range. One of the available approaches is based on performing several independent simulations at different λ values.⁹⁰⁻⁹¹ The number of simulations should be chosen such that the energy difference distributions at subsequent λ values overlap. This approach can be also applied in combination with other approaches, such as the importance sampling⁹⁰ and the adiabatic decoupling.⁹²

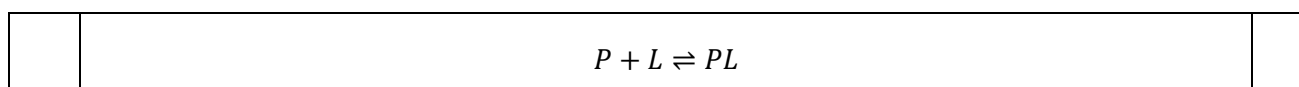
Once the relevant configurations have been sampled, the difference in free energy can be estimated. To this aim, different approaches are available.

Christ et al.⁸⁸ divided them in global and local methods. In the first category, approaches based on counting the number of times a given state is sampled and on exploiting energy differences are included. Local methods are those based on the computation of forces or transition probabilities. In the next paragraph, a global method based on energy differences (i.e. perturbation methods) is outlined. Then, thermodynamic integration is briefly presented as an example of local method to compute free energy differences. For a more detailed discussion on perturbation methods and TI, I refer to the work of Chipot and Pohorille.⁹³

2.8. Enhanced sampling methods in drug discovery

Thanks to the principles of statistical mechanics, quantitative estimates of important thermodynamic properties related to free energy can be obtained by molecular simulations. In the drug discovery field, the key thermodynamic quantity is the protein-ligand binding free energy, defined as the free energy difference between the unbound and bound states.^{27, 94}

Assume a two-state binding process, in which protein and ligand associate by non-covalent interactions.



This binding process will be characterized by the equilibrium association constant, K_a , expressed as the ratio between the concentration of the complex, PL , and of the dissociated species, P and L .

	$K_a = \frac{[PL]}{[P][L]}$	(2.118)
--	-----------------------------	---------

The reciprocal of the association constant defines the equilibrium dissociation constant, K_d , corresponding to the ligand concentration for which an equal probability of bound and unbound protein is achieved. Dealing with competitive inhibition assays, the inhibition constant, K_i , has the same chemical meaning of K_d .

Up to this point, we dealt with the Helmholtz free energy F referring to free energy. Let us now turn to the Gibbs free energy, G , which is the thermodynamic quantity experimentally determined at constant temperature and pressure.

Thus, assuming a two-state binding process at constant temperature and pressure, the equilibrium binding constants and the binding free energy are related by Equation 2.119:

	$\Delta G_{binding}^\circ = -k_B T \ln(K_a C^\circ) = k_B T \ln\left(\frac{K_d}{C^\circ}\right)$	(2.119)
--	--	---------

where C° is a constant defining the standard concentration. Since the standard binding free energy depends on the equilibrium constant and the reference value C° , this concentration has to be taken into account when a direct comparison between computed and experimental data is considered.⁹⁵ According to Eq. 2.116, standard binding free energy, $\Delta G_{binding}^\circ$, is considered a quantity directly related to the affinity of the ligand to the receptor.

As already mentioned, the determination of accurate free energy estimates is a challenging task. Dealing with large systems, such as protein-ligand complexes, the problem regarding insufficient sampling of the configuration space is particularly difficult. Thus, because of the exponential relationship in the configurational probability density in the canonical ensemble, $\rho_{NVT}(\mathbf{r}_i)$, it is evident that the configurations with higher potential energy, $U(\mathbf{r}_i)$, would be poorly sampled. In addition, if barriers larger than few $k_B T$ units need to be crossed, the sampling of the configuration space will be severely limited.

Equation 2.119 can be expressed as a function of the probability rate between bound and unbound states as:⁹⁶

	$\Delta G_{binding}^\circ = -k_B T \ln\left(\frac{\varphi}{\phi}\right) + k_B T \ln\left(\frac{C^{box}}{C^\circ}\right)$	(2.120)
--	--	---------

where $\varphi = \int_{\mathbf{r} \in bound} \rho(\mathbf{r}) d\mathbf{r}$ and $\phi = \int_{\mathbf{r} \in unbound} \rho(\mathbf{r}) d\mathbf{r}$.

The terms φ and ϕ relate to the probability of finding the protein in the bound and unbound states, respectively. The last term is a correction term to obtain the standard binding free energy, where C^{box} is the concentration of the interacting species included in the simulation box.

In principle, one needs to run long simulations to obtain a statistically robust probability ratio between bound and unbound states and a reliable estimate of binding free energy.

Molecular dynamics (MD) methods, together with Monte Carlo approaches, are the most common computational approaches to generate a Boltzmann distributed set of configurations, finding application in predicting and understanding the structure, function, and dynamics of interacting biomolecules. However, because of the energy barriers larger than few $k_B T$ units separating the associated and dissociated states in the conformation space, unbiased MD simulations are unable to properly recover transition rates. To overcome this limit, enhanced sampling procedures have been introduced to cross transition barriers retaining the correct probability distribution in the statistical ensemble.⁹⁷

Most of those enhanced simulation approaches rely on the fact that free energy is a state function. As a consequence, the free energy difference can be recovered regardless the path connecting the bound and unbound states. To this aim, computational methods to determine the free energy difference between two molecular states defining unphysical and physical paths have been proposed.⁹⁸ In practice, the calculation of the free energy difference follows a thermodynamic cycle, from which relative or “absolute” binding free energy can be computed.²⁷ Dealing with “absolute” free energy calculations, the free energy of a single state has to be known.

Accounting for relative binding free energy differences, one is interested in computing the difference in free energy between the same receptor in complex with two congeneric ligands, X and Y.

In this case, an unphysical path that smoothly and progressively transforms the bound and unbound states of ligand X into its congeneric ligand Y is defined. To this aim, a combined Hamiltonian is set up as a linear/non-linear interpolation of the potential energy functions describing the two molecular states as a function of the coupling parameter, λ . Thus, the unphysical path is sampled through a series of intermediate simulations covering the whole λ range. Perturbation methods or thermodynamic integration can then be applied to recover the free energy required to transform ligand X into its congeneric Y in the binding site, $\Delta G_{complex}^{X \rightarrow Y}$, and in the solution, $\Delta G_w^{X \rightarrow Y}$. The relative binding free energy is then obtained as the difference between those terms.

	$\Delta \Delta G_{binding} = \Delta G_{complex}^{X \rightarrow Y} - \Delta G_w^{X \rightarrow Y}$	(2.121)
--	---	---------

These calculations can be quite efficient if the two ligands are very similar to each other. If the ligands are different from a chemical point of view or if their bound states are separated by a high energy barrier,

problems related to the sampling of the conformation space can limit the applicability of those approaches. In addition, slow conformational changes occurring in response to the change of ligand need to be taken into account.⁹⁹

More complex thermodynamic cycles²⁷ can be applied to compute the “absolute” binding free energy difference between the dissociated and associated states of the protein-ligand complex. In these approaches, the ligand is reversibly changed into a fictitious non-interacting particle decoupled from its environment (bulk water and protein). In the simplest implementation, first the electrostatics is turned off, and then the van der Waals contributions are switched off. This approach, namely the double decoupling method (DDM) was first introduced by Jorgensen in 1988¹⁰⁰ and later formalized by other researchers.^{95, 101-103} Critical aspects of this procedure are related to the application of restraints, whose contribution to the free energy difference has to be taken into account. In addition, once the ligand is completely decoupled from the receptor, the binding site needs to be filled with water molecules. Some approaches have been proposed to overcome this problem.¹⁰⁴⁻¹⁰⁵ In 2016, Aldeghi et al.¹⁰⁶ applied this approach to a set of various inhibitors binding to bromodomain-containing protein 4 (BRD4).

Enhanced methods based on unphysical pathways are widely used to compute free energy differences. However, they are not able to give insights into the kinetics governing the formation of the binary complex. To address binding and unbinding kinetics, simulation methods relying on physical pathways are required to determine free energy barriers and possible intermediates along the path. In such approaches, the free energy can be reconstructed as a function of a reaction coordinate, which is known as the collective variable (CV), which takes into account the relevant (slow) degrees of freedom of the binding/unbinding event. Thus, from the projection of the free energy along the relevant degrees of freedom (i.e. the potential of mean force, PMF), both thermodynamic and kinetic quantities can be extracted.¹⁰⁷ Enhanced methods that explicitly use CVs act by biasing the molecular dynamics simulation along the predefined reaction coordinate. Usually the choice of the CVs to sample is not trivial requiring an extensive knowledge of the molecular system to be characterized in order to achieve a satisfying description of the underlying binding event.⁹⁷ Among the others, umbrella sampling⁹⁰ (US) is one of the most widely used equilibrium CV-based enhanced simulation method. In US, a harmonic restraining potential is applied to a series of windows equally spaced along a reaction coordinate. The PMF is then reconstructed via a reweighting procedure, such as WHAM.¹⁰⁸ Steered MD (SMD)¹⁰⁹⁻¹¹⁰ and metadynamics (MetaD)¹¹¹ are examples of the non-equilibrium methods. In SMD, a system can be brought from the initial to the final state applying a potential, usually imposing a constant velocity or force, on a subset of the system atoms along the predefined CV. The PMF can then be reconstructed and the free energy difference along the reaction coordinate can be derived. MetaD applies a biasing potential along the CV to allow the exploration of the configuration space. The bias is added as a sum of Gaussian functions deposited on visited points of the CV space inducing the system to adopt different and unexplored conformations. As a consequence, large energy barriers can be overcome and the CV space can be efficiently characterized. Variants of the MetaD have been proposed, such as the bias-exchange

metadynamics,¹¹² well-tempered MetaD,¹¹³ and a recent approach aimed to characterize kinetic properties from MetaD simulations.¹¹⁴

Enhanced methods that do not make explicit use of CVs are also available. For instance, one can bias the MD simulation heating all the system degrees of freedom (or part of them) at once or scaling the Hamiltonian. By definition, no a priori CV is required, as well as no knowledge about the rare events under investigation is needed. Because of that, those approaches are usually more broadly applicable in comparison to CV-based enhanced methods. Examples of computation approaches that fall into this category are temperature replica exchange MD (T-REMD)¹¹⁵ and Hamiltonian replica exchange MD (H-REMD)¹¹⁶. In general, replica exchange simulations aim at enhancing the sampling by running independent replica in slightly different ensembles, and periodically exchanging the coordinates of replicas between the ensembles. The acceptance ratio of the exchange moves relies on the Metropolis Monte Carlo algorithm. Then, thermodynamic quantities as a function of the temperature range can be recovered using multiple-histogram reweighting techniques.

Scaled MD¹¹⁷⁻¹¹⁸ is a recently proposed MD-based method, which was applied to unbinding kinetics predictions in a relative manner. It acts enhancing the transition between free energy minima by linearly scaling the system's potential energy. As a consequence, the rupture of all the fundamental physical interactions is facilitated, the probability of each microstate is altered, and the unbinding event can be observed in a reasonable computational time. At present, because of the peculiarities of the method, it is not difficult to recover the free energy difference between bound and unbound states directly from the biased simulation.

For detailed discussions about applications of enhanced methods to the (un)binding kinetics of drug-protein systems, I refer to the works of De Vivo,²⁷ Bernetti,¹¹⁹ and Bruce.¹²⁰

2.9. Methods to compute the absolute free energy: an overview

As outlined in the previous sections, the knowledge of the free energy difference between two molecular states allows the evaluations of fundamental thermodynamic properties. To accurately estimate differences in free energy, a unique probability distribution has to cover both states. To this aim, most of the computational approaches are based on the simulation of multiple unphysical intermediate states to improve the sampling of the overlap region. However, the high computational cost resulting from the sampling of the intermediate states does not allow the application of those approaches to real case studies.

This problem is circumvented dealing with the absolute free energies, F_A and F_B , which does not require any configurational overlap. The general computational strategy allows the evaluation of the absolute free energy by computing the free energy difference between the system of interest and a reference system for which the

free energy is known either numerically or analytically. Then, the system excess free energy is complemented by perturbation approaches. Common reference systems are the harmonic Einstein crystal¹²¹⁻¹²² for solids and the Lennard-Jones fluid¹²³ for liquids.

In 1990, Stoessel and Novak applied a harmonic solid in Cartesian coordinates to estimate the absolute free energies of small peptides in different conformations.¹²⁴ They presented the method as a tool to investigate not only low energy structures, but also higher energy (unstable) protein conformations. Knowing the absolute free energy of two molecular states, the computation of the difference in free energy is trivial. In this work, the main features of the reference system were pointed out, such as the availability of the reference free energy and the structural similarity between reference and system of interest.

Studies regarding the harmonic fluctuations about minimum-energy protein conformations lead back to the works of Go and Scheraga,¹²⁵ who computed the entropy of polypeptides undergoing harmonic fluctuations about their stable states. An extension of this approach was proposed by Karplus and Kushick in 1981.¹²⁶ They computed the covariance matrix of internal coordinates during MD rather than the normal mode analysis (NMA),¹²⁷ and then, by harmonically approximating the potential in internal variable space, they related local configuration entropy to fluctuations and correlations in internal variables. In addition, internal variables other than torsions were identified as those that mainly contribute to configurational entropy, suggesting the need to include other degrees of freedom in the energy characterization of protein conformations. In the same period, Hagler et al.¹²⁸ used the Monte Carlo method to determine relative vibrational free energies of different minimum-energy conformations confirming the role of local configurational entropy in relative conformational stability.

In the last 30 years, a considerable effort was spent trying to develop computational methods able to efficiently determine absolute free energies.

The hypothetical scanning (HS) method is a general approach that has been proposed by Meirovitch¹²⁹ to estimate the absolute entropy and free energy from Monte Carlo or molecular dynamics techniques. The HS method and the related local states method demonstrate that, like the energy, $\ln \rho$, where ρ is the Boltzmann probability density, can be seen in terms of product of transition probabilities among the system conformers. Thus, the absolute entropy and free energy can be obtained approximately from a given Boltzmann sample generated by Monte Carlo or MD simulations. The HS method has been applied to various systems, such as peptides in vacuum¹³⁰ and fluids.¹³¹⁻¹³²

In the late 1990s, Gilson et al.¹³³ suggested the mining minima approach to compute the conformational free energy of molecules. The approach is based on the identification of a set of low-energy conformations and on the evaluation of their contributions to the configurational integral by Monte Carlo techniques. The method has been applied to alanine oligopeptides and other small molecules.

Among the available computational schemes, the reference system method and the confinement approach represent two promising developments.

Ytreberg and Zuckerman¹³⁴ developed the reference system method addressing the problem in computing free energy arising from the lack of overlap between the end states. The method relies on the study made by Stoessel and Novak in 1990.¹²⁴ Ytreberg and Zuckerman defined a reference system characterized by high superimposition with the actual state eliminating the need of multistage approaches. The approach required the construction of normalized histograms of system's coordinates previously generated by finite-length simulations, and the generation of an ensemble of reference structures randomly chosen from the histograms. The reference energy and the potential energy from the force field needed to be computed for each structure in the reference ensemble. Then, free energy perturbation was applied to recover the absolute free energy of the system of interest.

In 2006, Tyka et al.¹³⁵⁻¹³⁶ presented the confinement approach taking inspiration from the previous works of Štrajbl and Warshel.¹³⁷⁻¹³⁸ Štrajbl and Warshel computed the entropic contribution to the activation energies of chemical reactions in which motions perpendicular to the reaction pathways were harmonically restrained. Then, they evaluated the free-energy cost associated to the presence of restraints by free energy perturbation. The confinement method is a path-independent method relying on the transformation of each state into a harmonic (solid) reference state in which the vibrational entropy is the sole component of the total entropy, allowing the direct computation of the relative free energy. Tyka et al. successfully applied the method to compute the side-chain entropies of a β -hairpin forming peptide in a variety of backbone conformations in implicit solvent.

In 2007, Tyka et al.¹³⁶ suggested an extension of the confinement method to compute absolute free energies of fluids under periodic boundary conditions. In this application, the liquid system is transformed into a harmonic, solid, reference state. The occurrence of the liquid-solid phase transition during the confinement made difficult to accurately estimate the system free energy. In addition, dealing with diffusive systems, the inadequacy of the harmonic reference state for fluid systems became clear. To overcome this problem, Tyka et al. developed an approach based on the reassignment of the reference positions of individual atoms during the simulations. This approach was explored in my computation, and for this reason it is briefly discussed in Chapter 4.

In 2009, Karplus et al.¹³⁹ improved the efficiency of the confinement method showing the results on two biological systems, alanine dipeptide in vacuum and β -hairpin from protein G in implicit solvent. Similarly to the original version of the method, the absolute free energy of the conformers was determined by progressively restraining the molecular conformations to harmonic basins, whose free energy was determined by normal mode analysis. Note that the presence of implicit solvation overcomes the problems related to the diffusion of bulk molecules and to the presence of the first-order phase transition. Karplus et al.

identified possible source of errors in the confinement calculations and proposed ways to correct them. They also tested the dependence of the calculated free energy values on the structure-based definition of molecular states used to extract conformations from the corresponding basins. Finally, they tested the performance of a different approach based on the covariance matrix instead of the harmonic approximation.

In 2015, Esque and Cecchini¹⁴⁰ extended the applicability of the Karplus' version of the confinement method to analyze the difference in free energy of explicitly solvated conformers. This variant of the method, referred to as confinement/solvation free energy (CSF) method, aims at computing the free energy of a molecular system in explicit solvent by transforming the actual conformation in solution into the harmonic reference state in vacuum, whose free energy is computed by normal mode analysis.

The free energy of the conformer in explicit solvent is then computed as a correction of the harmonic free energy in vacuum including the system an-harmonicity and the solvent contribution. Those terms are computed by the confinement of the solute only and the hydration free energy of the confined conformer via free energy perturbation (FEP), respectively. The approach was tested on three peptides: the alanine dipeptide, the met-enkephalin, and the deca-alanine. It provided accurate free energy estimates as well as insight on the conformational stability of biomolecules by separating the contributions from the an-harmonicity (entropy) and the solvent.

In 2003, Henchman has proposed a different approach based on cell theory to characterize pure liquid systems.¹⁴¹ In the original cell method,¹⁴² each molecule is treated as moving in separate "cells". Thus, the energy of the system can be decomposed into individual effective potentials, each of them dependent only on a single molecule. The free energy follows the partition function, which is obtained by integrating over the effective potentials. The original cell method was not able to accurately model liquids and to carry out the decomposition of the energy of the system for highly correlated systems. In the first development of the method proposed by Henchman,¹⁴¹ the effective potential chosen to approximate the system potential energy surface is a three-dimensional isotropic harmonic function, parametrized based on the average potential energy per atom and the average magnitude of the force on each atom. The partition function is then computed from these effective potentials to derive all the thermodynamic properties of interest over a range of temperatures and densities. The method was applied to liquid argon as test system.

In 2007, Henchman extended his method to liquid water,¹⁴³ modeling each molecule as a rigid body moving in a six dimensional anisotropic harmonic potential. Seven quantities were measured from a single equilibrium simulation of water to retrieve free energy, enthalpy, and entropy. The free energy of water and its excess free energy were evaluated by the partition function of the quantized harmonic oscillators. The entropy of water was decomposed into translational, rotational, and conformational terms. In this application, the harmonic potential emerged as a model suitable to interpret the liquid-phase behavior.

The approach was further validated in 2008, when Stennett and Henchman tested the method to compute the classical and quantum free energy of liquid water and ice Ih.¹⁴⁴ At the same time, Zielkiewicz verified the general applicability of the harmonic approximation to retrieve the entropy of liquid water reproducing the calculations of Henchman in a more rigorous way.¹⁴⁵

More recently, Habershon and Manolopoulos¹⁴⁶ addressed the problem of calculating the classical free energies of liquids and solids described by molecular models with intra-molecular flexibility. In this work, they focused on the computation of the classical phase diagram of an empirical water model that possesses intra-molecular flexibility, namely q-TIP4P/F.¹⁴⁷ Applied to the solid state, their approach relies on a step of thermodynamic integration to compute the difference in free energy between the real system and a Debye crystal reference state with anisotropic harmonic force constants derived from the force field. Regarding the liquid phase, they applied the same computational approach replacing the Debye crystal with a Lennard-Jones (LJ) fluid, whose excess free energy has been previously tabulated over a range of state-points making it an ideal reference for free energy calculations of the liquid state.¹²³ However, unlike the LJ fluid, liquid water is a molecular fluid. Thus, to deal with this problem, they estimated the intra-molecular contribution to the free energy by thermodynamic integration between the real system and a fluid reference system, whose interaction sites were fixed at the molecular center-of mass. Then, the absolute free energy is computed including two additional terms. One represents the free energy of a system of N molecules which interact solely via LJ sites located at the center of mass of each molecule. The second term includes the contribution of each water molecule that can freely rotate and vibrate about its center of mass under the action of its intra-molecular potential. The approach presented by Habershon and Manolopoulos can be in principle be generalized to any other molecular system with intra-molecular flexibility. In the Chapter 4 of this thesis, an approach relying on the Debye reference state is presented, validated, and tested to estimate hydration free energies of organic molecules.

Finally, the very recent work of Li, Totton and Frenkel is mentioned.¹⁴⁸ They proposed a “black-box” computational scheme as a generic approach to compute chemical potentials/free energies in dense liquids. The method relies on the fact that the chemical potentials of the solubilized and crystallize phase of a solute coincide at the equilibrium in saturated solutions. The main limitation of the practical applicability of this principle is the accurate estimate of the chemical potentials of both the crystalline and solution phases. Thus, they developed an approach based on standard alchemical free energy methods, such as thermodynamic integration and free energy perturbation, consisting of two parts. The first part requires the systematic extension of the Einstein crystal method to compute the absolute solid free energies for each intermediate crystalline stage at arbitrary temperature and pressure. Then, in the second part, a flexible cavity method is applied to yield accurate estimates of the excess solvation free energies. In this approach a repulsive cavity is first created in the solvent before inserting the solute in a series of steps. Thus, the method enables convenient solubility estimates of general molecular crystals at arbitrary thermodynamic conditions where solid and solution coexist. As first application of the method, the solubility of OPLS-AA-based naphthalene

solvated in SPC water was studied obtaining results in good agreement with experimental data at various temperatures and pressures.

One year later, they extended the application of the method computing the solubility of a set of soluble organic/inorganic compounds, including phenantrene, calcite, aragonite, and caffeine.¹⁴⁹

3. Understanding protein-ligand unbinding kinetics in kinases through electrostatics-driven adiabatic bias molecular dynamics

3.1. Aim of the project

Equilibrium binding metrics, such as thermodynamic binding affinity, sometimes cannot satisfactorily describe what happens under non-equilibrium conditions of living organisms. In those cases, a key determinant of *in vivo* pharmacological response is the lifetime of the binary drug-target complex, expressed as the reciprocal of the dissociation rate constant, k_{off} (i.e. drug-target residence time, $1/k_{\text{off}}$).²⁸ Tuning drug-target residence time directly impacts the duration of the therapeutic effect and the *in vivo* drug efficacy. It also influences other pharmaceutically relevant drug properties, such as selectivity¹⁵⁰ and safety.¹⁵¹ Thus, the increased difference in residence time between primary and secondary targets can cause a compound to kinetically select one receptor over another. Modulating drug selectivity thus affects the related therapeutic index.²⁴

Experimental techniques can directly estimate kinetic rate constants but miss in mechanistic atomic level details. When rationally designing the kinetics of drug binding/unbinding, it can be interesting to get insights into the particular molecular determinants of the highest free energy barrier, such as drug-target interactions, protein conformational states, ligand flexibility, and water dynamics.¹⁵²

In principle, molecular dynamics (MD)-based methods can be applied to evaluate kinetic rate constants. However, they are limited by the need to extensively sample the conformational space characterized by free energy barriers, with dissociation times far larger than the timescales usually sampled by computer simulations. One solution is to combine MD simulations with Markov State Model (MSM)¹⁵³ analysis. This strategy was adopted by Buch et al. to define a kinetic model for the two-state binding process of trypsin-benzamidine complex.¹⁵⁴ They identified multiple intermediate states and estimated the standard free energy of binding and absolute association and dissociation kinetic rates. However, MSMs are designed to obtain rates from pre-sampled MD data. They do not address the main problem with MD, which is how to sample states that are high in free energy.

In recent years, several enhanced sampling methods have been proposed to overcome the sampling problem of MD simulations.^{28, 119-120} Whilst these computational techniques are usually used to reconstruct the free energy of binding, they can also be used to compute absolute and relative kinetic rate constants of series of ligands. In this context, metadynamics (META-D)¹¹¹ is a well-known and widely used non-equilibrium enhanced sampling method, which can accelerate the exploration of configurational space by using Gaussian potentials to bias the dynamics of the system along particular collective variables (CVs). Although

metadynamics was originally designed to characterize static properties, Tiwary and Parrinello slightly modified the original method to obtain transition rates between metastable states.¹⁵⁵ Recently, Casasnovas¹⁵⁶ applied this variant to a pharmaceutically relevant study of the unbinding kinetics of a representative inhibitor of p38 MAP kinase. A good estimate of the absolute dissociation rate constant, k_{off} , was obtained, and the unbinding mechanism and rate-limiting steps were completely characterized with the MSM derived from state-to-state metadynamics simulations.

Other researchers¹⁵⁷⁻¹⁵⁸ used a different approach has been used to predicting unbinding kinetics. Here, the goal is the ranking, rather than the absolute residence time values. Recently, metadynamics-based protocols were used to compute relative dissociation rate constants. Callegari used a classical implementation of metadynamics to prioritize a series of cyclin-dependent kinase 8 (CDK8) inhibitors. Bortolato provided a generally applicable computational protocol based on the combination of adiabatic bias molecular dynamics¹⁵⁹ (ABMD) and well-tempered metadynamics¹⁶⁰ (*wt*-META-D) to simulate protein-ligand unbinding events. A peculiar scoring function allowed the chemical series to be classified based on computed residence times, while an atomic solvation factor gave insights into the water dynamics during ligand dissociation. Also recently, the tau random accelerated molecular dynamics (tau-RAMD) was successfully used to rank HSP90 inhibitors.¹⁶¹ Mollica et al. ranked series of ligands with scaled molecular dynamics simulations (scaled MD).¹¹⁷ Scaled MD was applied to several pharmaceutically relevant cases¹¹⁷⁻¹¹⁸ providing residence time-based ranking correlations that were in good agreement with experimental kinetic data. The relatively modest requirements of computational resources and the wide applicability of the protocol make it suitable for the hit-to-lead and lead-optimization phases of drug discovery. Scaled MD promotes the transition between energy minima by linearly scaling the potential energy by an arbitrary factor, λ . By scaling the potential, the probability of each microstate is altered, the rupture of interactions stabilizing protein-ligand complexes is facilitated, and unbinding events are observed in shorter timescales. Moreover, the absence of a peculiar reaction coordinate makes the protocol quite general and easy. Harmonic restraints are applied to preserve the correct protein folding, leaving the regions involved in the unbinding process unrestrained. Thus, although a collective variable is not required, this kind of simulations still requires a priori information, in that one must define a binding region for the application of the restraints. Due to the use of scaled potentials, the mechanistic and energetic details of the explored unbinding processes are lost or at least heavily approximated. That is, the scaled MD protocol is very effective but also quite “brutal” in obtaining the unbinding process.

Against this background, we here report a fast, efficient, and widely applicable computational protocol for simulating protein-ligand dissociation events. The method is based on adiabatic-bias molecular dynamics (ABMD) coupled with an electrostatics-driven collective variable. A key feature is that it delivers atomic-resolution mechanistic information about dissociation pathways. The new approach was applied to two pharmaceutically relevant kinases: Glucokinase (GK) and Glycogen Synthase Kinase 3 beta (GSK-3 β). GK is a cytoplasmic enzyme with a recognized role in the maintenance of glucose homeostasis. It is a molecular

target for type 2 diabetes.¹⁶² GSK-3 β is a proline-directed serine/threonine kinase involved in Alzheimer's disease.¹⁶³ In this section, the flexibility of the approach in terms of applicability to various chemotypes and its predictive power are shown. In both cases, the predicted unbinding time-based ranking correlations were in good agreement with the experimental data. The first case was retrospective, whereas the second was prospective and subsequently experimentally validated through SPR experiments. New crystallographic structures were also determined to provide appropriate starting configurations for estimating the residence times. The present methodology is fast enough for use in the hit-to-lead and lead-optimization phases, while offering a level of accuracy to meaningfully discriminate congeneric and non-congeneric chemical series in terms of residence time. Additionally, being the adiabatic, it provides mechanistic insights into the drug-target dissociation mechanisms.

3.2. Computational methods

3.2.1. Electrostatics-driven adiabatic bias molecular dynamics (elABMD)

3.2.1.1. Adiabatic bias molecular dynamics (ABMD)

Adiabatic bias molecular dynamics (ABMD) is a simulation technique that generates MD trajectories connecting points in conformational space separated by activation barriers. This biased MD methodology was first used to simulate the crystallization in amorphous solids. It was subsequently applied to the unfolding of lysozyme¹⁵⁹ and of different fibronectin type 3 domains.¹⁶⁴

To use ABMD, one must define the starting state and final state of the system, as well as a specific reaction coordinate to which a time-dependent harmonic biasing potential is applied. By specifying the extremal points to be joined it is possible to monitor whether or not the system evolves spontaneously toward the final target state. Therefore, the biasing potential can be added to the potential energy function only when the system attempts to move in the opposite direction with respect to the desired final end point. ABMD is thus particularly interesting because it gives a realistic description of the evolution of a system to an external perturbation, leaving its short-time dynamics relatively unperturbed.

The additive perturbation of the potential energy function $U(\rho(t))$ has the form of a pawl-and-ratchet system:

	$U(\rho(t)) = \begin{cases} 0.5K(\rho(t) - \rho_m(t))^2 & \text{if } \rho(t) > \rho_m(t) \\ 0 & \text{if } \rho(t) \leq \rho_m(t) \end{cases}$	(3.1)
--	--	-------

where:

	$\rho(t) = (CV_t - CV_0)^2$	(3.2)
	$\rho_m(t) = \min_{0 \leq \tau \leq t} (\rho(\tau) + \eta(t))$	(3.3)

$\rho_m(t)$ represents the minimum value of $\rho(\tau)$ observed so far during the simulation, which is defined as the square difference between the instantaneous reaction coordinate value, CV_t , and its target value, CV_0 . $\eta(t)$ represents an optional noise term, set to 0 in the present study.

If, in the simulation step from t to $t + \Delta t$, $\rho(t)$ spontaneously decreases with respect to $\rho_m(t)$ (i.e. $\rho(t) \leq \rho_m(t)$), the external perturbation is zero, $\rho_m(t)$ is updated (i.e. $\rho_m(t)$ is set equal to $\rho(t + \Delta t)$), and $U(\rho(t))$ is modified accordingly. The time-dependent harmonic bias potential is applied to prevent $\rho(t)$ from increasing significantly, inducing $\rho(t)$ to stay in the last visited $\rho_m(t)$. The main peculiarity of this sampling method is that thermal motions due to the finite temperature rule the progression along the reaction coordinate in terms of visited states.

The value of the force constant, K , affects the magnitude of the backward fluctuations of the reaction coordinate, the speed of the system's evolution toward the final state, and the adiabatic character of the transformation. In general, by an appropriate choice of the force constant, the perturbation due to $U(\rho(t))$ can be kept small compared to the system's total energy. Therefore, the system will move on an almost constant energy surface and the transformation will be adiabatic.

In this work, the spring constant, K , was empirically defined based on the desired speed/accuracy trade-off. We tested different values until we observed a satisfying discrimination of compounds included in the series. As a general rule, we fixed a priori a maximal amount of simulation time per replica, and then we devised a proper spring constant to allow the discrimination of the compounds under investigation.

3.2.1.2. Electrostatics-driven collective variable (eCV)

Electrostatic contributions represent the natural long range forces that drive molecular recognition. In particular, electrostatic interactions between charged entities affect binding specificity¹⁶⁵ and association/dissociation rates.¹⁵² In the present work, we chose the electrostatics variable reported in Spitaleri et al.¹⁶⁶ The peculiarity of this variable is that it does not use the original system charges, but rather apply equal and formal charges to both the ligand and the binding site. Formally, let L be the set of ligands atoms and S the set of pockets atoms, the collective variable can be expressed as the sum of terms reported in Equation 4:

	$CV(t) = \sum_{i \in L, s \in S}^n \frac{Q_i Q_s}{r_{i,s}(t)} \exp(-r_{i,s}(t)/\lambda)$	(3.4)
--	--	-------

where, r are the inter-atomic distances, $\exp(-r/\lambda)$ is a decay function aimed at avoiding unnecessary long ranged forces, and λ is the parameter that rules the spatial range of the decay (heuristically set to 10 Å). Fictitious charges are therefore positioned on every atom, in particular $Q_L = Q_S = 1$ for all the atoms.

This variable has been found to give collective and smooth behaviors over time.¹⁶⁶ We did not consider any tuning of the formal charges and all have the same absolute values and are equal in sign. The combination of eCV with ABMD, namely eLABMD, leverages the “gentleness” of ABMD and the “smoothness” of the electrostatic field lines to facilitate the unbinding process, leaving the system free of restraints. In the present application, the target value eCV_0 (Eq. 3.2) was trivially set to zero to formalize the absence of protein-ligand electrostatic interactions. By definition, the electrostatic bias due to eCV is spread over protein and ligand repulsing atoms, on which fictitious charges with same sign are positioned. Ligand atoms were selected as the biggest common atoms subset in the series, following a geometrical conservation criterion. In our case, 5 and 12 atoms were identified in GK and GSK-3 β series, respectively. Protein residues were those located within 6 Å of the ligand. The repulsive GK residues included in the selection were 57-65, 92-94, 155, 206, 208, 210, 211, 214, 216-218, 231, 248, 447-449, and 451-454, according to the work of Mollica et al.¹¹⁸ The repulsive GSK-3 β residues were identified with the help of Nanoshaper¹⁶⁷ and were 62, 67, 70, 72, 83, 85, 110, 132-138, 141, 185, 186, 188, 199, and 200 according to PDB codes 4ACG, 4ACD, and 4ACH.

3.2.2. Simulation Setup and Analysis

To assess the reliability of the method, we first compared the results generated by scaled molecular dynamics (sMD) simulations on the Glucokinase (GK) complexes in Mollica et al.¹¹⁸ with those obtained by eLABMD. Another aim of this validation was also to enrich the previous sMD analysis with mechanistic details. For all the simulations, we used the same structures as those reported in Ref. 118.

Subsequently, a series of Glycogen synthase kinase 3 beta (GSK-3 β) ATP-competitive inhibitors was studied. In Table 3.2, chemical structures are depicted with their corresponding numbering conventions. For GSK-3 β , crystal structures in complex with **1**, **2**, and **3**¹⁶⁸ were retrieved from the Protein Data Bank (PDB IDs: 4ACG, 4ACD, and 4ACH, respectively). In the present work, the pyridine core of the crystalized ligand **1** (PDB ID: 4ACG) was converted in a pyrazine ring by Maestro from Schrödinger (Version 11.1.012, Release 2017-1)¹⁶⁹ without modifying the geometrical orientation of the bound state. This structural modification was necessary to make **1** consistent with the overall series of pyrazine derivatives. When experimental protein-ligand complexes were not available, namely for **4**, **5**, and **6**, new crystal structures were generated (PDB IDs: 6HK4, 6HK3, and 6HK7, respectively). The protonation state of all GSK-3 β

inhibitors was assigned at pH 7.0 using the Molecular Discovery MoKa software package v2.6.6.¹⁷⁰ Two positive charges were assigned on nitrogen atoms located on the terminal 4-methylpiperazine and pyrrolidine groups of **1** (60.1%). One positive charge was positioned on the nitrogen of the 4-methylpiperazine group of **2** (83.4%), **3** (84.8%), **5** (84.9%), and **6** (84.8%). Zero charge was assigned to the morpholine analogue of **2**, namely **4** (95.8%). The ligand bonded parameters were assigned by Antechamber/GAFF force field.⁶³ The point charges were obtained from RESP¹⁷¹ calculations via the NwChem software.¹⁷² The GSK-3 β receptor was prepared with the Protein Preparation Wizard in Maestro. The co-crystallized ligand tautomers were selected by starting from MoKa predictions. All crystallographic water molecules were deleted except to the one located in the internal ATP-binding pocket of complexes with **1** and **2** (i.e. HOH2107 and HOH2097, respectively). PROPKA¹⁷³ included in the Protein Preparation Wizard was used to predict the protonation state of protein residues at pH 7.3. All systems were prepared using the BiKi Life Sciences 1.3 software package.¹⁷⁴ For proteins and ions, the Amber ff14⁶² was chosen. Water was described by the TIP3P¹⁷⁵ model. The input structures were placed in the box center, whose margins were located at 12 Å from the complex. The electroneutrality was preserved by adding sodium/chloride ions as needed. Short-range electrostatic interactions were treated with the Verlet cut-off scheme, and long-range ones with the Particle Mesh Ewald (PME) method.⁷⁴ In both cases, the cut-off was fixed at 11 Å. Periodic boundary conditions (PBC) were applied to avoid boundary effects by finite size, and to approximate the infinite system by the simulated one. Energy minimization was carried out with 5,000 steepest descent steps. The MD leap frog integrator was chosen to run the systems. All bonds were constrained during equilibration. The equilibration protocol comprises the following steps: the system was thermalized at 300 K in three steps using the Bussi-Parrinello thermostat,⁷⁹ for a total of 0.3 ns of dynamics. Subsequently, 1 ns of dynamics was performed in the NPT ensemble until the average pressure of the system was equilibrated to 1 atm according to Parrinello-Rahman barostat.⁷⁵ The heavy atoms of the protein backbone were restrained with 1,000 kJ mol⁻¹ nm⁻¹ force during the first and second equilibration steps. The final GK and GSK-3 β systems contained approximately 80,000 and 60,000 atoms, respectively. All MD simulations were performed using Gromacs 4.6.1¹⁷⁶ patched with the BiKi Life Sciences version of the Plumed 2.0¹⁷⁷ beta plugin to run adiabatic bias molecular dynamics (ABMD).¹⁵⁹ All the simulations were performed in the NVT ensemble setting 2 fs as time step. Velocities were randomly initiated before each production run. The protein's structure was left free of restraints. We run a statistics of 10 independent replicas of 10 ns for each GK complex on one GPU workstation, running approximately at 15 ns/day. A statistics of 20 independent replicas of 30 ns each was performed for each GSK-3 β complex running on the Linux cluster Marconi at CINECA national supercomputing center.

Analysis of the simulations was performed using the BiKi Life Sciences 1.3 software package.¹⁷⁴ The unbinding time was computed as the time when the ligand was surrounded by a water shell of 3 Å for the first time.

The water shell radius was selected comparing the Pearson and Spearman coefficients resulting from the unbinding time based ranking correlations computed for the series of GK, fixing the water shell radius at 2.5

Å, 3 Å, and 3.5 Å. Varying the water shell, the predictions of intermediate compounds in terms of unbinding time were affected. The water shell radius giving the best compromise between Pearson and Spearman coefficients was selected (Table 3.1).

Table 3.1. Pearson and Spearman coefficients at different water shell radius^a

Water shell radius	Pearson coefficient	Spearman coefficient
2.5	0.67	0.71
3.0	0.71	0.75
3.5	0.58	0.75

^a Water shell radius is expressed in Å.

Bootstrap analysis¹⁷⁸ was performed to establish whether an increased number of simulations were needed, and to define robust mean estimations, on which reliable ranking correlations were built. The Hydra analysis module included in BiKi Life Sciences 1.3 software package¹⁷⁴ was used to evaluate the distribution of water molecules around the GSK-3β binding site when complexed with structurally different ATP-competitive inhibitors (i.e. **5** and **7**). The analysis of water molecules positioning is reported in the Results section.

3.3. Experimental methods

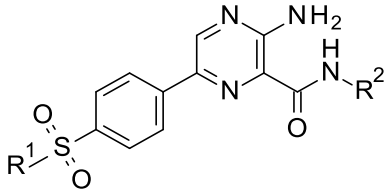
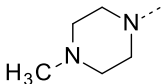
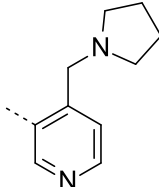
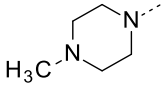

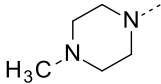
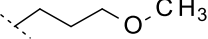
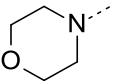

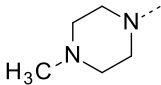
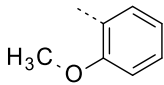
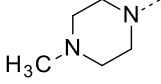
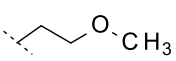
For the sake of completeness, the experimental methods have been included in this thesis. Valentina Piretti expressed, purified, and crystallized the kinase GSK-3β. She also performed the kinetic assays of the GSK-3β inhibitors. Rita M. C. Di Martino synthesized and purified the GSK-3β inhibitors. Shailesh K. Tripathi refined the GSK-3β crystal structures.

3.3.1. Chemistry

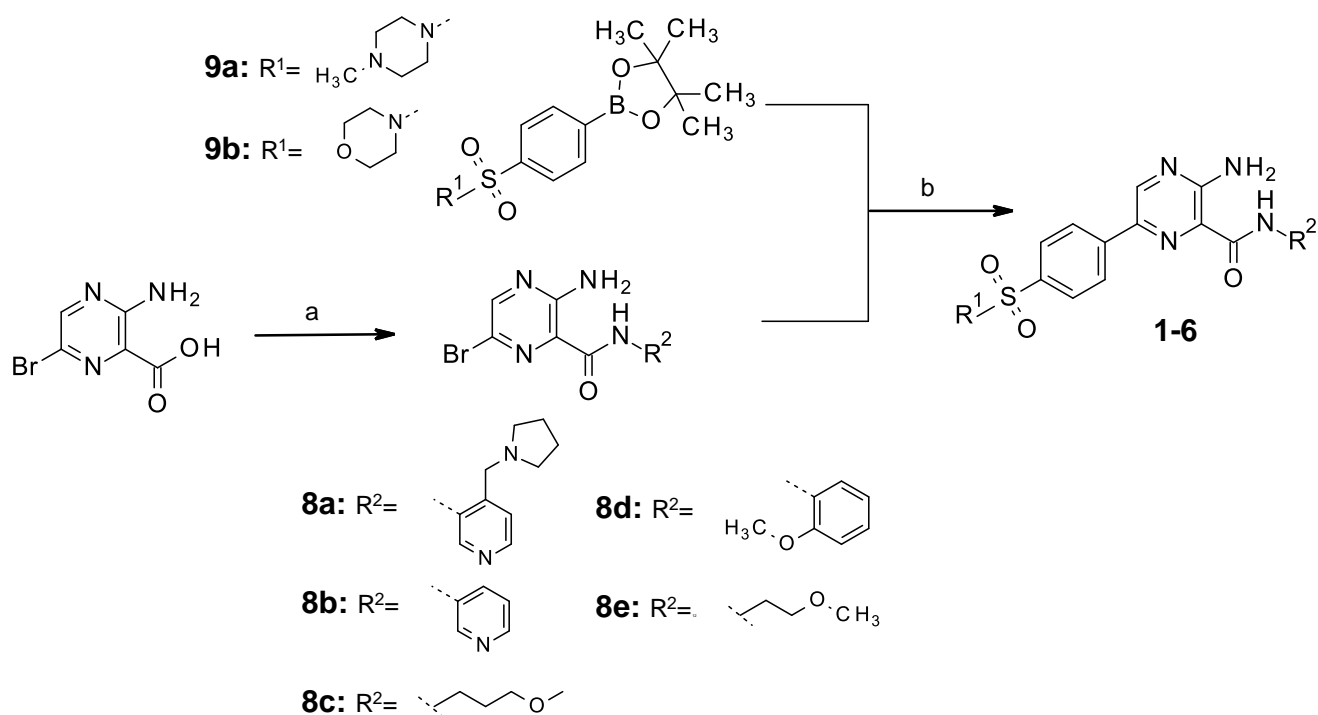
The selected GSK-3β inhibitors **1-6** were prepared by applying a general synthetic strategy, which was adapted to that reported by Berg et al.¹⁶⁸ A Suzuki cross-coupling reaction between the appropriate phenylboronic acid pinacol ester (**9a** or **9b**) and the apposite pyrazine bromide (**8a-e**) gave the desired sulfonamides **1-6** with moderate to very good yields as described in Scheme 3.1. 6-bromopyrazines (**8a-e**) were, in turn, prepared by applying a reaction of direct amide formation from 3-amino-6-bromo-pyrazine-2-carboxylic acid and the proper alkyl- or arylamine in the presence of 1-[bis(dimethylamino)methylene]-1*H*-1,2,3-triazolo[4,5-*b*]pyridinium 3-oxid hexafluorophosphate (HATU) as coupling agent and N,N-diisopropylethylamine (DIPEA) as base (Scheme 3.1). Moreover, a Miyaura borylation reaction between bis(pinacolato)diboron and the commercially available 4-((4-bromophenyl)sulfonyl)morpholine afforded the

intermediate phenylboronic acid pinacol ester **9b** (Scheme 3.2). Amine **11** required for the synthesis of **8a** was synthesized via a two-step procedure reported in Scheme 3.3. A reductive amination reaction of N-Boc-protected 3-aminopyridine-4-carbaldehyde with pyrrolidine in the presence of sodium triacetoxyborohydride (STAB) as reducing agent gave **10**, which was finally converted in **11** through a reaction of N-Boc-deprotection using trifluoroacetic acid (TFA) in dichloromethane.

Table 3.2. Chemical structure of synthesized compounds 1-6.

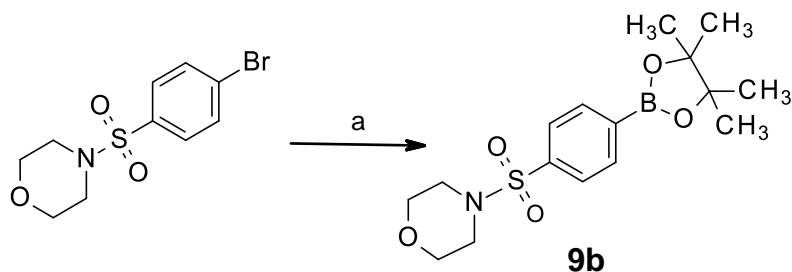
		
Cpd	R ¹	R ²
1		
2		
3		
4		
5		
6		

Scheme 3.1. General synthetic procedure for compounds 1-6 and intermediates 8a-e^a



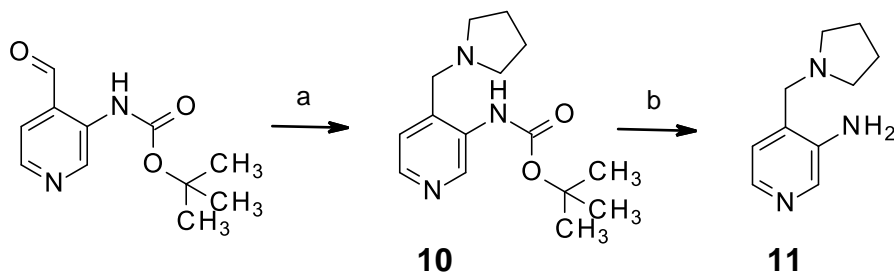
^a Reagents and conditions: (a) alkyl- or arylamine, HATU, DIPEA, dry DMF, rt, 40 min-12 h, 34-100 %; (b) PdCl₂(dppf)·CH₂Cl₂, Na₂CO₃ (aq, 2 M), toluene/EtOH (5:1), 80 °C, Ar, 2-22 h, 41-81 %.

Scheme 3.2. Miyaura borylation reaction: synthesis of intermediate 9b^a



^a Reagents and conditions: (a) bis(pinacolato)diboron, CH₃COOK, PdCl₂(dppf)·CH₂Cl₂, dry 1,4-dioxane, 80 °C, Ar, 3 h, 68 %.

Scheme 3.3. The two-step synthesis of intermediate amine 11: reductive amination and N-Boc deprotection reactions^a



^a Reagents and conditions: (a) pyrrolidine, CH₃COOH, STAB, 1,2-dichloroethane, rt, Ar, 24 h, 100 %; (b) TFA, CH₂Cl₂, 0 °C to rt, 24 h, 100 %.

3.3.2. GSK-3 β expression and purification

The DNA sequence of human GSK-3 β full length (1-420) was cloned in pFB-LIC-Bse vector (kindly provided by Dr. Opher Gileadi, SGC-Oxford). The resulting transfer vector was sequence verified and transformed into *E.coli* DH10Bac cells to obtain the recombinant bacmid-DNA. Sf9 cells (Expression Systems LLC, Davies - USA) were seeded in a six-well plate at 1.5×10^6 cells/well in ESF-921TM medium (Expression systems) and transfection was performed using FuGENE[®] HT reagent (Promega). Plate was incubated at 27°C and recombinant baculovirus was harvested 55 hours post-transfection (P0 stock). The high titer virus stock (P2) was generated by two rounds of amplification and used for protein expression. High Five (H5) cells (Expression Systems LLC, Davies - USA) were infected with P2 stock at an initial density of 1.5×10^6 cells/ml in ESF-921TM medium, incubated at 27°C and harvested by centrifugation 72 hours after infection. Cells were resuspended in lysis buffer (20mM Tris-HCl pH 8.0, 0.5M NaCl, 10mM imidazole, 5% glycerol, 5mM MgCl₂, 1mM DTT and Complete protease inhibitors (Roche)) and lysed by sonication at 50% amplitude (10'' pulse, 20'' pause on ice, total time 2 minutes). Lysate was incubated with Benzonase[®] (Merck-Millipore) for 10 minutes at room temperature and clarified by centrifugation for 1 hour at 30000xg, 4°C. Supernatant was incubated with equilibrated NiNTA agarose resin (Qiagen) on gentle rotation for 1 hour at 4°C. The flow-through was collected by centrifugation at 500xg for 15 minutes and the NiNTA resin loaded on a polypropylene column (Bio-Rad). The column was washed with binding buffer (20mM Tris-HCl pH 8.0, 0.5M NaCl, 10mM imidazole, 5% glycerol, 1mM DTT) and eluted with 0.3M imidazole in binding buffer. Collected fractions were diluted 10-fold in Buffer A (20mM Hepes pH 7.5, 5% glycerol, 1mM DTT) and loaded onto a HiTrap SP HP 1ml column (GE Healthcare), eluted by a step gradient of Buffer B (20mM Hepes pH 7.5, 1M NaCl, 5% glycerol, 1mM DTT). Phosphorylated GSK-3 β (pTyr216) eluted at 100mM NaCl and was used for SPR experiments. All protein molecular weights and phosphorylation states were confirmed by LC mass spectrometry. Purified protein was stored at -80°C in 20mM Hepes pH 7.5, 150mM NaCl, 5% glycerol, 1mM DTT.

3.3.3. Kinetic characterization of GSK-3 β inhibitors by Surface Plasmon Resonance (SPR)

SPR measurements were carried out on a Pioneer FE (Pall FortéBio) at 25°C. GSK-3 β was immobilized on a HisCap SensorChip (Pall FortéBio), and then let equilibrate in immobilization buffer (20 mM Hepes pH 7.5, 150 mM NaCl, 0.01% Tween20). To improve the chip surface stability, an approach based on the capture-coupling protocol proposed by Rich et al.¹⁷⁹ was employed. The method is a hybrid of capture and amine coupling chemistry, where the His tag serves to pre-concentrate and orient the protein onto the surface for subsequent covalent crosslinking via the activated carboxyl groups. The SensorChip surface was activated with 500 μ M of NiCl₂ in immobilization buffer at a flow rate of 10 μ L/min. After the activation of the surface, GSK-3 β diluted in immobilization buffer at 7.5 μ g/mL was immobilized by injecting 100 μ L at a flow rate of 10 μ L/min. The surface was stabilized by amine coupling with the injection of 7.5 μ L of 0.2 M

NHS/0.4 M EDC diluted 1:10 in H₂O at a flow rate of 15 μ L/min, followed by an injection of 35 μ L of 1.0 M ethanolamine pH 8.0 at a flow rate of 5 μ L/min to block unreacted groups. The chip surface was cleaned from Ni²⁺ ions by injection of 60 μ L EDTA in immobilization buffer at a flow rate of 20 μ L/min. The covalent coupling procedure resulted in a stable baseline. GSK-3 β was immobilized alternatively on Flow cell (FC) 1 or FC 3, whereas FC 2 was used as reference. Typical immobilization levels ranged from 3500 to 4500 RU. Binding experiments were performed in binding buffer (50 mM Tris-HCl pH 7.5, 250 mM NaCl, 0.01% Tween20) supplemented with 5% DMSO. Tested compounds were solubilized in 100% DMSO and then diluted in binding buffer by serial doubling. The top concentration for each compound was optimized to improve the accuracy of the steady-state and kinetic dissociation constants, K_Ds. In binding assays, a flow rate of 40 μ L/min was set up. Referring to the numbering of Table 3.2, compounds **3**, **5**, and **6** did not require a regeneration step. For these compounds, association was measured for 5 minutes, and dissociation for up to 20 minutes depending on peculiar off-rates. Due to high off-rates, compounds **1**, **2**, and **4** required a regeneration step. Association and dissociation were recorded for 3.5 and 2 minutes, respectively. An optimized regeneration solution, i.e., 15% DMSO in H₂O, was injected at a flow rate of 30 μ L/min for 2 minutes to allow the complete dissociation of tight protein-ligand complexes.

3.3.4. Analysis of binding data

All data analysis and processing was performed using Pioneer Qdat Software (Pall FortéBio). Binding Response was recorded in real time as a change in surface plasmon resonance measured in resonance units (RUs). The equilibrium analysis was performed plotting the responses against the analyte concentration and fitting the data to a 1:1 binding isotherm. Kinetic curves were fit to the Simple model of Qdat Software including bulk refractive index offsets. Binding analysis of all compounds injected across immobilized GSK-3 β was replicated two times on two different sensor chips with the same capacity. Although fitting data from multiple capacity surfaces provides more information about binding reactions, in most cases enough binding information is available from a single surface.¹⁸⁰ K_D values obtained from equilibrium and kinetic fit agree if saturation in the experiment is achieved. The difference between K_D values derived from both fits (Δ K_D) was used as an indicator to assess the quality of the experiments (Table 3.3). For each compound, only the replicate showing the lowest Δ K_D was considered to define the unbinding time-based ranking correlations reported in Table 3.7. The corresponding affinity and binding curves are reported in the Appendix.

Table 3.3. Selection of two replicates of surface plasmon resonance (SPR) binding experiments^a

Cpd	K_D st-st	K_D kin	ΔK_D	k_{on}	k_{off}	$t_{r,exp}$
1	8.6 ± 0.2	1.31 ± 0.02	7.29	$1.005 \pm 0.004 \text{ E+06}$	$1.32 \pm 0.02 \text{ E-03}$	757.58
	2.79 ± 0.04	1.64 ± 0.01	1.15	$1.001 \pm 0.003 \text{ E+06}$	$1.64 \pm 0.01 \text{ E-03}$	609.76
2	14.5 ± 0.1	4.73 ± 0.01	9.77	$6.78 \pm 0.01 \text{ E+05}$	$3.21 \pm 0.01 \text{ E-03}$	311.53
	8.4 ± 0.1	4.29 ± 0.02	4.11	$8.97 \pm 0.03 \text{ E+05}$	$3.85 \pm 0.02 \text{ E-03}$	259.74
3	17.75 ± 0.09	10.90 ± 0.02	6.85	$8.19 \pm 0.01 \text{ E+05}$	$8.925 \pm 0.006 \text{ E-03}$	112.04
	20.8 ± 0.1	10.80 ± 0.03	10.00	$8.22 \pm 0.02 \text{ E+05}$	$8.88 \pm 0.01 \text{ E-03}$	112.61
4	4.9 ± 0.1	2.98 ± 0.01	1.92	$1.203 \pm 0.003 \text{ E+06}$	$3.59 \pm 0.01 \text{ E-03}$	278.55
	2.81 ± 0.04	1.53 ± 0.01	1.28	$1.667 \pm 0.005 \text{ E+06}$	$2.55 \pm 0.02 \text{ E-03}$	392.16
5	31.4 ± 0.6	18.93 ± 0.05	12.47	$2.411 \pm 0.006 \text{ E+05}$	$4.565 \pm 0.005 \text{ E-03}$	219.06
	44.2 ± 0.6	14.77 ± 0.03	29.43	$1.867 \pm 0.003 \text{ E+05}$	$2.758 \pm 0.002 \text{ E-03}$	362.58
6	45.9 ± 0.3	42.3 ± 0.2	3.6	$1.286 \pm 0.005 \text{ E+06}$	$5.4400 \pm 0.0001 \text{ E-02}$	18.38
	111 ± 1	56.1 ± 0.3	54.9	$1.155 \pm 0.006 \text{ E+06}$	$6.4800 \pm 0.0002 \text{ E-02}$	15.43

^a Both K_D steady-state and K_D kinetics are expressed in nM, k_{on} in $s^{-1} M^{-1}$, k_{off} in s^{-1} , experimental residence times, $t_{r,exp} = 1/k_{off}$, in s. The errors reported refer to the quality of the model fit.

3.3.5. Crystallization of GSK-3 β in complex with compounds 4-6

Crystallization trials of GSK-3 β were conducted using Hampton Research PEG/Ion Screen I and II. Optimizations of PEG3350 concentration, buffer and pH were carried out to achieve well diffracting crystals. Co-crystals of GSK-3 β with compounds **4** or **5** were obtained by sitting drop method mixing 0.1 μ l of protein-ligand solution with 0.1 μ l reservoir (**5**) or 0.5 μ l of protein-ligand solution with 0.5 μ l reservoir (**4**) and equilibrating against 80 μ l of precipitant solution at 20 °C. Prior crystallization each protein-ligand solution was obtained adding 600 μ M compound to a solution containing 200 μ M GSK-3 β in 20 mM HEPES pH 7.5, 150 mM NaCl, 20 mM DTT, 5% glycerol (w/v), and incubating at room temperature for two hours. The protein-ligand solution was then mixed in the sitting drop with a reservoir solution containing 8% tacsimate pH 7.0 and 20% PEG3350 for compound **4** and 0.1 M Bis-Tris pH 6.5, 2% tacsimate pH 6.0 and 20% PEG3350 for compound **5**. Single crystals appeared within 4 days.

GSK-3 β crystals in complex with **6** were obtained by sitting drop and subsequent soaking. Protein crystals were grown at 20 °C by mixing 0.1 μ l of 150 μ M GSK-3 β , 20 mM HEPES pH 7.5, 150 mM NaCl, 20 mM

DTT, 5% glycerol, 1 mM AMP-PNP, 2 mM MgCl₂ with 0.1 µl precipitant solution containing 0.2 M KF and 22% PEG3350. Resulting crystals were soaked for two hours in 1.2 mM C50 in 0.2 M KF and 22% PEG3350.

Crystals were cryo-protected in mother liquor with 20-25% glycerol and 3X molar excess of inhibitor and flash-frozen in liquid nitrogen prior to data collection.

3.3.6. Data collection and structure determination

Diffraction data was collected at beamline XRD1 of Elettra synchrotron, Trieste, Italy. Crystals corresponding to ligand **4** and **5** diffracted to 2.2 and 2.34 Å respectively, while ligand **6** diffracted to 3.0 Å. Integration of reflection files was performed using XDS.¹⁸¹⁻¹⁸² Integrated data was scaled using AIMLESS¹⁸³ in CCP4 suit.¹⁸⁴ Phasing was performed by molecular replacement using phaser.¹⁸⁵ Previously published GSK-3β structure (PDB ID: 4ACD) was used as initial search model in phaser. Structures were refined using phenix.refine.¹⁸⁶ Model optimization and editing were performed in Coot.¹⁸⁷ In the case of data corresponding to ligand **6**, refinement was performed using low resolution refinement in REFMAC,¹⁸⁸⁻¹⁸⁹ wherein the restraints were generated using PDB model 4ACD in proSMART. Diffraction data corresponding to complex structures of ligand **4** and **5** was relatively better compared to **6**. Molecular replacement placed two protein molecules in the asymmetric unit in P1 space group for structures corresponding to ligand **4** and **5**. In case of ligand **6**, a single molecule was placed in the asymmetric unit in C222₁ space group. Although each crystal varied in terms of quality of data, unambiguous electron density was observed corresponding to each ligand along with the residues surrounding the ligand.

Structures corresponding to ligand **4**, **5**, and **6** were deposited in the Protein Data Bank (PDB IDs: 6HK4, 6HK3, and 6HK7, respectively).

3.4. Results

3.4.1. A retrospective validation of elABMD protocol: the GK case

3.4.1.1. Validation of elABMD protocol

In the retrospective phase, the proposed elABMD-based protocol was applied to the allosteric GK activators (GKAs) that Mollica et al.¹¹⁸ characterized with scaled molecular dynamics (sMD) coherently with the experimental k_{off} . The chemical series included compounds displaying T-shaped and linear geometries (Fig. 3.1). In that paper the obtained results were in line with the main hypothesis of the original sMD method,¹¹⁷ according to which the success of unbinding time predictions is related to the chemical similarity of the compounds.

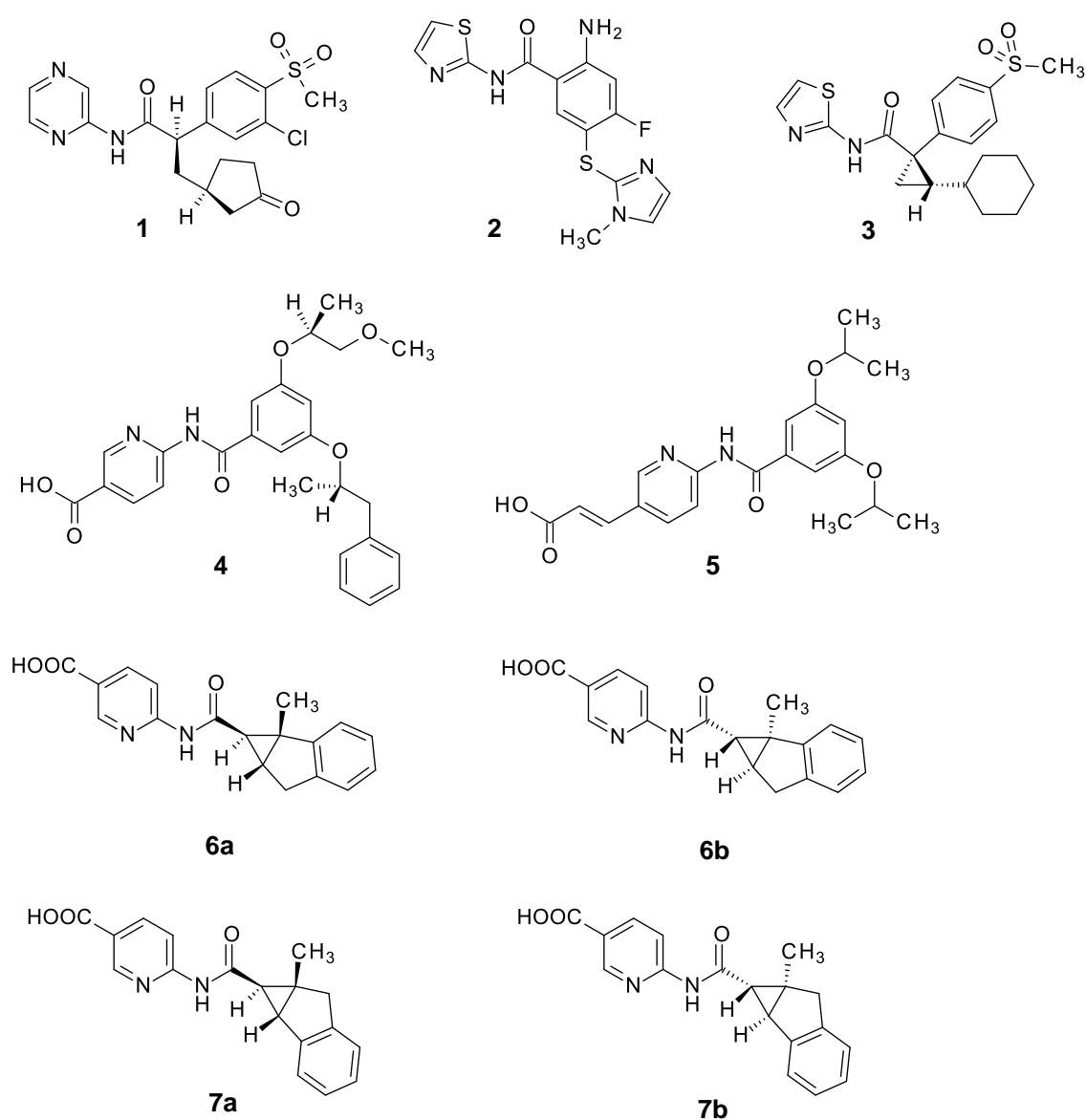


Figure 3.1. Chemical structures of the GK activators (1-7).

We initially identified the system-dependent force constant, K , by running some exploratory unbinding simulations of the slowest and fastest compounds of the series varying K . By setting the force constant equal to $2.0\text{E-}15 \text{ [kJ/mol]}^{-3}$, we collected ten simulations of 10 ns for each GKAs complex satisfactorily differentiating the compounds included in the series. The enantiomers of the two racemic mixtures (i.e. ligands **6_{a,b}** and **7_{a,b}**) were individually simulated, and relative unbinding times were estimated averaging the results obtained for both compounds. To assess the statistical robustness of our observations and to establish whether an increased number of replicas was needed, a bootstrap analysis of mean unbinding times was carried out. Bootstrapped estimations of eLABMD unbinding times were then used to build the predicted unbinding time-based ranking correlations, which were in good agreement with those obtained by experimental residence times (i.e. $t_r = 1/k_{\text{off}}$) and sMD results with scaling factor, $\lambda = 0.5$ (Table 3.4 and Fig. 3.2). When comparing correlations obtained with sMD and eLABMD, both techniques clearly distinguished the tightest and weakest ligands (i.e. **1** and the racemic mixtures **6_{a,b}** and **7_{a,b}**). eLABMD underestimated **5** with respect to the experimental data, although correctly ranked. **4** was overestimated by sMD and in particular by eLABMD that ranked **4** as the tightest compound of the series, raising doubts about the reliability of experimental k_{off} or the ligand parameterization.

Table 3.4. Experimental residence time ($t_{r,\text{exp}}$), scaled MD ($t_{r,\text{mean}}$), and eLABMD ($t_{r,\text{mean}}$) unbinding times for each compound of GK series^a

Cpd	<u>Experimental</u>		<u>Scaled MD</u>		<u>eLABMD</u>	
	$t_{r,\text{exp}}$	Rank $t_{r,\text{exp}}$	$t_{r,\text{scaled}}$	Rank $t_{r,\text{scaled}}$	$t_{r,\text{mean}}$	Rank $t_{r,\text{mean}}$
1	8.3	1	105.1 ± 10.1	1	6.57 ± 0.76	2
2	2.3	4	29.3 ± 5.3	5	5.62 ± 1.06	5
3	2.7	3	38.9 ± 7.1	4	5.91 ± 1.14	3
4	1.6	5	92.9 ± 7.3	3	7.96 ± 0.86	1
5	6.3	2	99.7 ± 6.7	2	5.88 ± 0.86	4
6 _{a,b}	0.7	6	25.9 ± 3.9	6	5.26 ± 1.09	6
7 _{a,b}	0.2	7	24.7 ± 3.0	7	2.89 ± 0.63	7

^a Experimental residence times, $t_{r,\text{exp}} = 1/k_{\text{off}}$, are expressed in s; sMD and eLABMD unbinding times (i.e., $t_{r,\text{mean}}$) are expressed in ns. The sMD and eLABMD predictions and relative estimations of error are obtained via a bootstrap procedure. Spearman coefficients for sMD and mean eLABMD unbinding time-based ranking correlations were computed equal to 0.89 and 0.61, respectively.

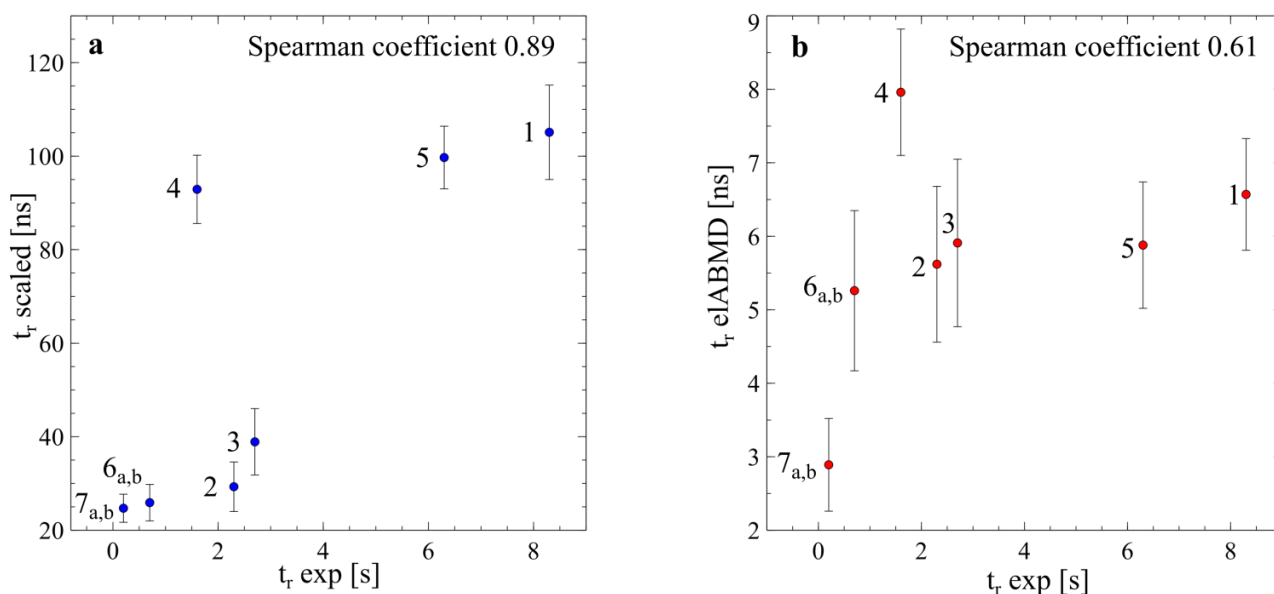


Figure 3.2. Experimental versus computational residence time. The rankings obtained with (a) sMD and (b) the bootstrapped estimations of mean elABMD unbinding time are reported.

elABMD predicted similar unbinding times for **5**, **3**, **2** and the racemic mixtures of enantiomers **6_{a,b}**, despite different experimental residence times. To improve the differentiation of these compounds, we increased the accuracy of unbinding simulations by tuning down the harmonic bias strength and thus increasing the simulation time. To identify the most suitable high-resolution force constant, we started by halving the K value applied in the first screening and doubling the simulation time. We found $K = 1.0\text{E-}15 \text{ [kJ/mol]}^{-3}$ and 30 ns of maximal simulation time be the best compromise for our system. **2** was compared to the enantiomer **6_a**, whose structure was considered representative of both racemic mixtures. Table 3.5 reports the bootstrapped estimations of mean unbinding times collected from both statistics. By decreasing the harmonic bias strength and increasing the resolution of unbinding simulations, we succeed in ranking **5/3** and **2/6_a** in agreement with experimental residence time. That is, the resolution (i.e. the accuracy of elABMD unbinding simulations) could be tuned easily by modulating the harmonic bias strength.

Table 3.5. High resolution unbinding time predictions using bootstrap estimations of mean unbinding times^a

Cpd	<u>Exp</u> $t_{r,\text{exp}}$	<u>elABMD</u>			
		$K = 2.0\text{E-}15$		$K = 1.0\text{E-}15$	
		$t_{r,\text{mean}}$	rate	$t_{r,\text{mean}}$	rate
5	6.3	5.88 ± 0.86	1.00	23.71 ± 3.51	0.70
3	2.7	5.91 ± 1.14		19.61 ± 4.44	

Cpd	<u>Exp</u>	<u>elABMD</u>			
		K = 2.0E-15		K = 1.0E-15	
		t_{r,mean}	rate	t_{r,mean}	rate
2	2.3	5.62 ± 1.06	0.87	21.21 ± 4.70	0.61
6 _a	0.7	4.93 ± 1.07		12.98 ± 4.32	

^a Experimental residence times (t_{r,exp}) are expressed in s; elABMD unbinding times (t_{r,mean}) are expressed in ns; the force constant, K, is expressed in [kJ/mol]⁻³. The rate between the lower and higher bootstrapped estimations of mean unbinding time is reported to highlight the differentiation between the dissociation times after increasing the accuracy of unbinding simulations.

Therefore, elABMD-based protocol differentiated structurally different compounds, prioritizing the most promising in terms of residence time. In particular, elABMD ranked the overall GKA series in acceptable agreement with experimental data, associating longer unbinding times with T-shaped ligands (i.e. **1**, **4**, **5**, **3**), and shorter unbinding times with those displaying a more linear geometry (i.e. **2** and enantiomers **6_{a,b}** and **7_{a,b}**), confirming previous computational predictions.

By considering the bootstrap estimations of median unbinding times, we obtained a similar unbinding time-based ranking correlation (see the Appendix).

3.4.1.2. Unbinding path analysis and Structure-Kinetic Relationships (SKRs)

During the dissociation phase, a high number of different unbinding paths were obtained. Each of them is characterized by a peculiar energy profile and different occurrence probabilities. To address the complexity of ligand unbinding kinetics, we investigated the possible relationship between the ligand structure and the explored unbinding paths. We did this by comparing the dissociation mechanisms of structurally similar compounds that led to complete ligand solvation at different unbinding times.

We started from the observation¹¹⁸ that GKAs displaying linear scaffolds exhibited faster experimental and computational off-rates than T-shaped ligands. Several factors were suggested to influence the off-rates of linear GKAs, including the limited number of rotatable bonds, the absence of the stabilizing network created by the polar sulphonyl group, the presence of chemical groups displaying greater water affinity to replace the hydrophobic moieties of branched compounds, and the rather linear shape of GK allosteric pocket. Referring to the numbering of GK in complex with **1** (PDB ID: 4NO7), it has been noticed that the displacement of Tyr215 side chain was needed to accommodate compounds characterized by T-shaped geometry into the allosteric pocket suggesting an induced fit binding mechanism and slower off-rates.¹¹⁸ Part of GK allosteric site is formed by the highly flexible region comprising residues spanning from Leu47 to Gly68, namely the connecting loop 1 (CL1)¹⁹⁰⁻¹⁹¹, which connects the small and large lobes of the enzyme (Fig. 3.3 and 3.4). It

was reported that GKAs efficacy in increasing GK responsiveness to glucose should be related to the capacity of ligands to fix the conformation adopted by CL1 when the enzyme is in its active (i.e. closed) conformation.

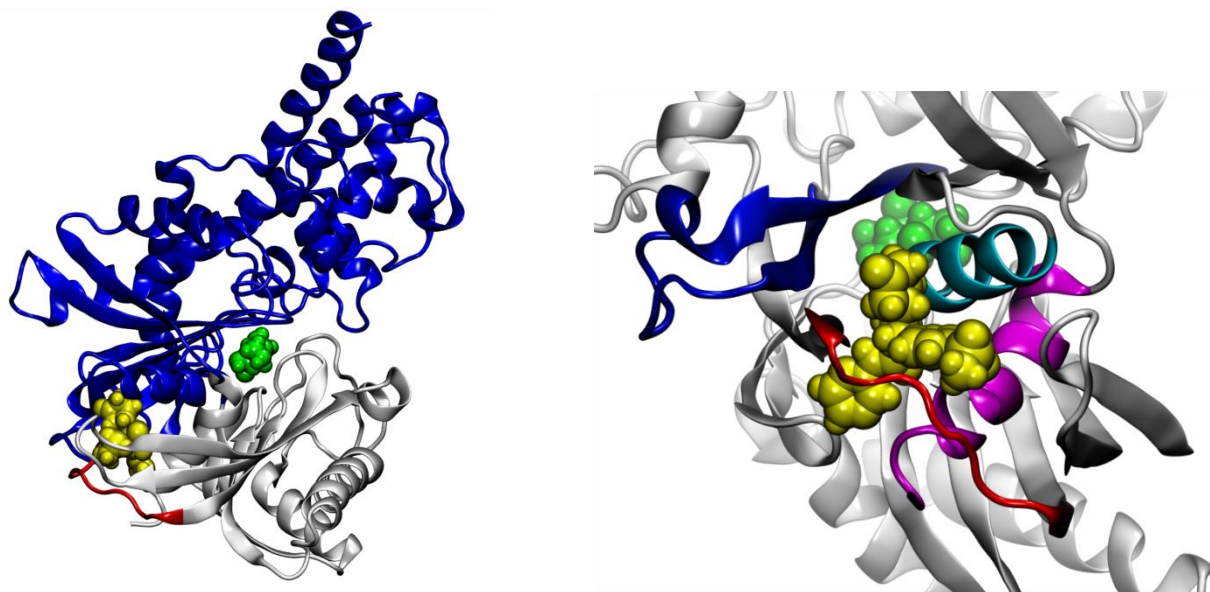


Figure 3.3. Structure of Glucokinase (GK). (Left) Active conformation of GK in complex with glucose (green) and the allosteric activator **1** (yellow). Large and small lobes are reported in blue and white, respectively. The 64-72 loop facing the allosteric site is red colored. (Right) Allosteric binding site of GK. The 64-72 loop (red), $\alpha 5$ helix (cyan), $\alpha 13$ helix (magenta), $\beta 10$ and $\beta 11$ sheets (blue) are reported. Glucose and ligand **1** are green and yellow colored, respectively.

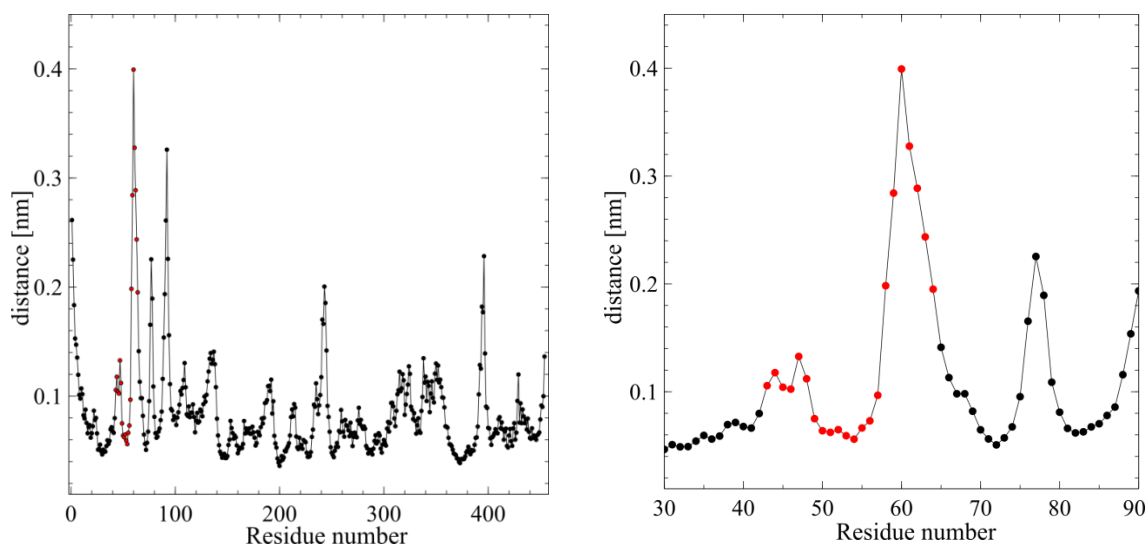


Figure 3.4. (Left) Root mean square fluctuations against the residue number of a representative unbinding trajectory explored by ligand **1**. (Right) The residues included in the most flexible region of GK corresponding to the connecting loop **1** are red colored.

T-shaped structured GKAs: ligands **1** and **3**

We first considered **1** and **3**, which are characterized by small T-shaped geometry and slow experimental and computational off-rates. The destabilization of T-shaped ligand complexes started with the rupture of the hydrogen bonds with Arg63 sidechain. The aliphatic ring and the substituted phenyl ring slightly arranged inside the binding pocket while the highly flexible 64-72 loop changed conformation. We observed that **1** and **3** explored one recurring unbinding path at different unbinding times, transiting over the 64-72 loop with respect to the bound state orientation of the ligand.

Faster unbinding events occurred when 64-72 loop adopted a conformation stabilized by Arg63 sidechain interacting with the negatively charged Asp158 and Asp160. The wide protein arrangement left enough space to the pyrazine and thiazole moieties of **1** and **3**, respectively, to move towards the solvent resulting in the dissociation from the allosteric pocket (Fig. 3.5, Left). When 64-72 loop maintained the bound state conformation, the same dissociation mechanisms led to slower unbinding paths hampered by the rather linear shape of the allosteric GK pocket and the non-linear geometry of **1** and **3**. In these cases, the unbinding process was slowed down or hindered by the insertion of heterocyclic moieties into the transient protein cavity formed by residues 60, 62, 156, 159, 201, 203, 207, 211, 452. The rotation of Val62 sidechain drove the aromatic rings to move inside the pocket. The direct interaction between the carbonyl oxygen of the aliphatic ring of **1** and Arg250 sidechain contributed to the slowing of the ligand dissociation.

Occasionally the 64-72 loop adopted a different conformation, which was stabilized by transient hydrogen bonds involving Tyr61, Arg67, Ser69, Gln98, and the adjacent 241-250 loop (Fig. 3.5, Right). In **1** bound state, Pro66 conformation allowed Glu67 sidechain to be oriented towards Hie218, creating an additional stabilizing interaction. The hydrogen bond between Glu67 and Hie218 was missing in GK in complex with **3**. The steric hindrance created by the alternative protein arrangement induced **1** and **3** to find a different unbinding path transiting under the 64-72 loop. This resulted in slow dissociation mechanisms hampered by the unfavorable geometries of both ligands and binding pocket. Before reaching complete solvation, **1** and **3** transiently interacted with protein portions surrounding the allosteric binding pocket, such as the $\alpha 5$ helix (205-215), the adjacent loop connecting $\beta 10$ - $\beta 11$ sheets (241-250), and $\beta 3$ - $\beta 4$ (94-97).

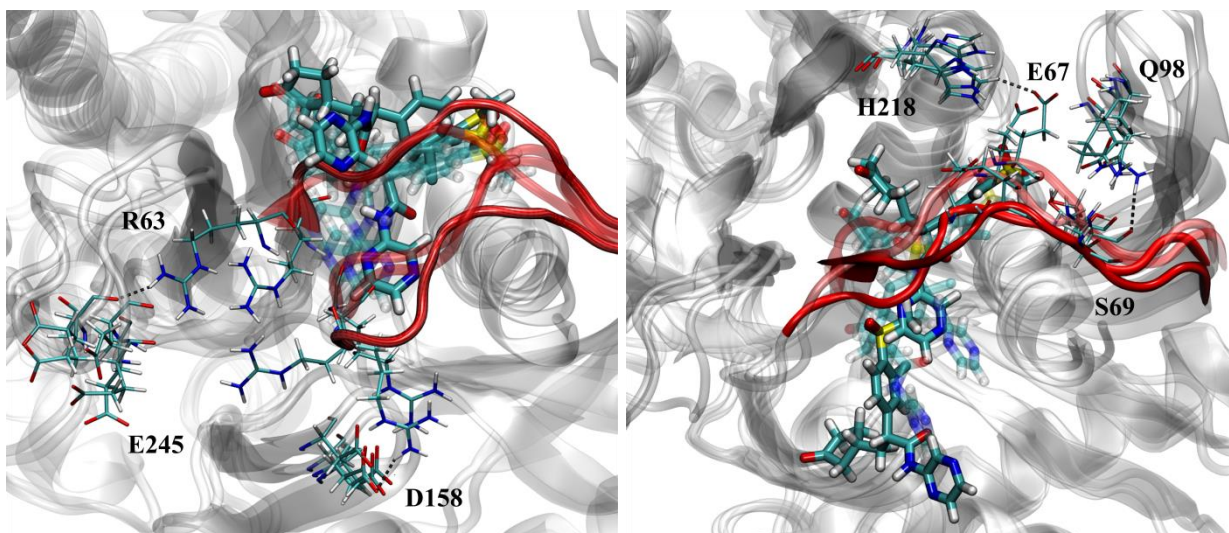


Figure 3.5. Progression of conformational rearrangements of 64-72 loop during two representative unbinding trajectories of **1**. (Left) Fast unbinding path: **1** transits over the CL1. (Right) Slow unbinding path: **1** transits under CL1. The 64-72 loop is red colored. The key protein residues involved in the conformational changes are represented in stick.

Linear structured GKAs: ligand **2** and enantiomers **6_{a,b}**, **7_{a,b}**

We then analyzed compounds displaying a more pronounced linear shape and faster off-rates, namely **2** and the racemic mixtures of **6_{a,b}**, **7_{a,b}**. Looking at the dissociation mechanisms explored by linear GKAs, we identified two main unbinding paths at different unbinding times. These differed in that the transition occurred under/over the 64-72 loop.

2 was involved in fast dissociation mechanisms when the imidazole moiety rotations inside the allosteric binding pocket induced a slight displacement of the allosteric loop, which opened a gap adjacent to β 10 and β 11 sheets of the large lobe. This structural rearrangement allowed the progressive hydration of **2**, which rapidly dissociated transiting over the 64-72 loop (Fig. 3.6, Left). When **2** was sterically locked into the binding site and the allosteric loop maintained its bound state conformation, we observed **2** slowly dissociating under the 64-72 loop at different unbinding times depending on Arg63 and Lys459 sidechains fluctuations (Fig. 3.6, Right).

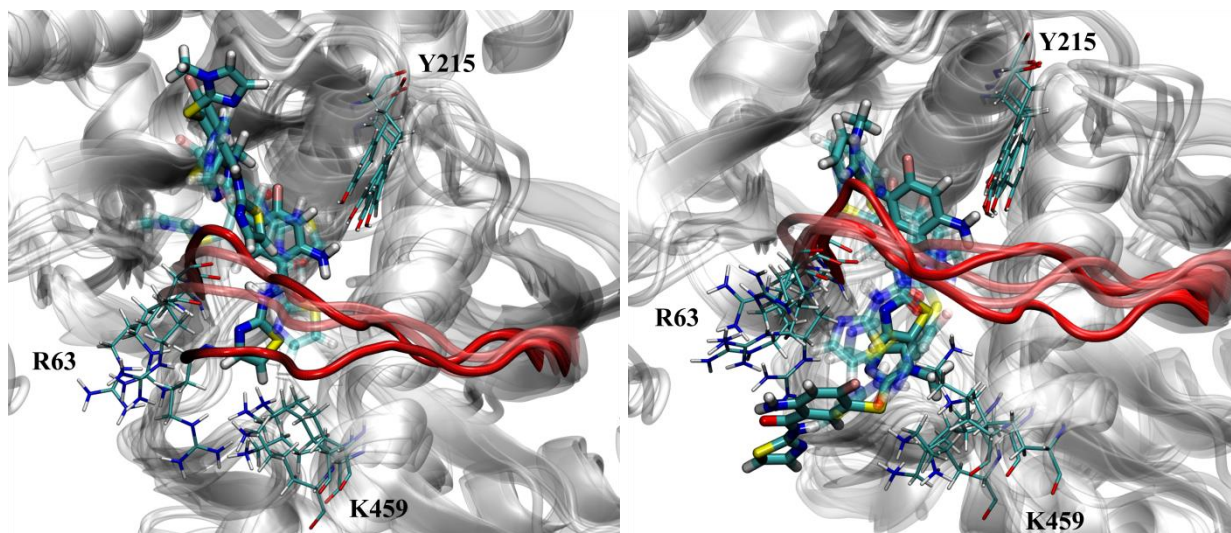


Figure 3.6. Fast (Left) and slow (Right) representative unbinding paths explored by **2**. The 64-72 loop is red colored. The protein residues involved in the dissociation mechanisms are represented in stick.

Referring to the racemic mixtures of enantiomers, we observed different unbinding behaviors. Rotations of the rigid aromatic moiety started the dissociation mechanism, causing the rupture of the π - π stacking interactions with Tyr214 sidechain and the loss of the key hydrogen bonds with Arg63.

Faster dissociation events occurred when the 64-72 loop adopted conformations that were transiently stabilized by residue-residue interactions (e.g. Glu67-Hie218). When Arg63 and Lys459 sidechains changed orientation, a gap under the 64-72 loop opened, and linear enantiomers rapidly dissociated from the binding site (Fig. 3.7, Left). Less frequently, we observed that different rotation degrees of the rigid aromatic moiety and fluctuations of the allosteric loop, induced **6_a** and **7_a** to fast dissociate transiting over the 64-72 loop (Fig. 3.7, Right). Interestingly, the same unbinding path was explored when the transition under the allosteric loop was hindered by Arg63 and Lys459 sidechains pointing the carbonyl oxygen of **6_{a,b}** and **7_{a,b}**. This interactions network delayed wide conformational rearrangements of the allosteric loop and promoted the establishment of one hydrogen bond between Lys459 sidechain and the solvent exposed carboxylic oxygen, increasing the relative unbinding time.

After the dissociation from the allosteric pocket, enantiomers transiently interacted with the protein regions surrounding the allosteric pocket, such as those including α 5 helix (205-215) and loop connecting α 5 helix to β 9 sheet (216-221).

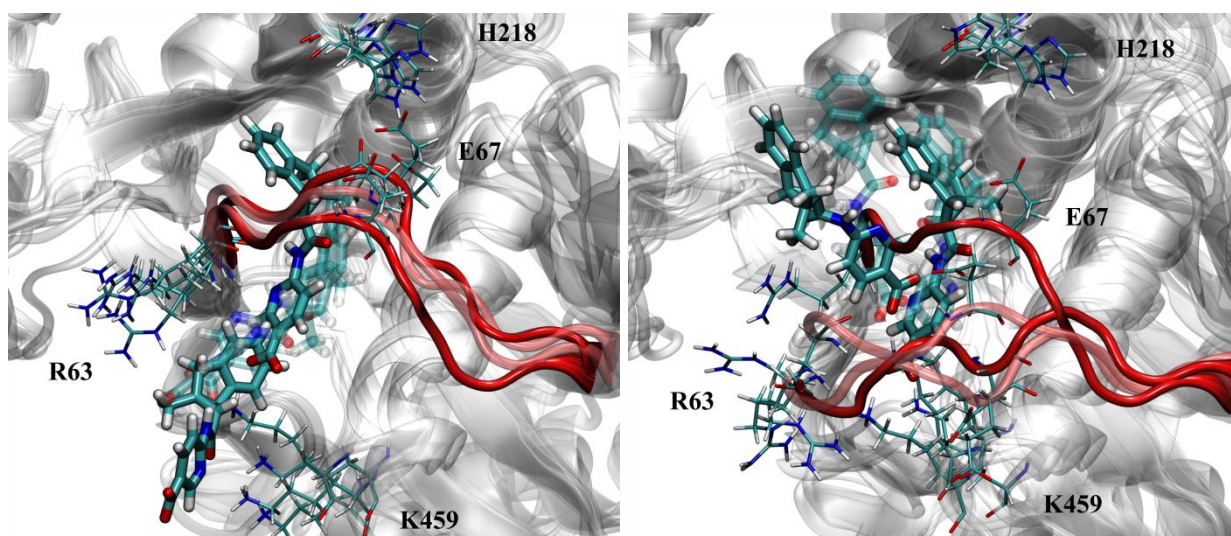


Figure 3.7. Fast (Left) and slow (Right) representative unbinding paths explored by the **7_a**. The 64-72 loop is red colored. The protein residues involved in the dissociation mechanisms are represented in stick.

Ligands **4** and **5**

Finally, we focused on **4** and **5**, whose larger structures displayed characteristics of both branched and linear GKAs. These characteristics included bearing electron rich oxygen atoms in the same region of the sulphonyl group of **1** and **3**, and carboxyl groups in a similar position to linear enantiomers. Moreover, the larger size of **4** and **5** was suggested to increase the enthalpic contribution to the overall unbinding activation energy extending the predicted unbinding times.

The bound states of **4** and **5** were mainly stabilized by Arg63 backbone, which directly interacted with the amide moiety and the electron donor atoms of the adjacent heterocycle, similarly to T-shaped and linear compounds. Moreover, the stability of the complex was increased by the hydrophobic surrounding of the substituted phenyl moiety formed by Val62, Pro66, Ile211, Tyr214, Tyr215, Met235, Leu451, Val452, and the transient interactions, which were established by the electron-rich oxygen atoms included in the flexible substituents of the phenyl ring.

Looking at the trajectories collected for **4** and **5**, we identified one recurring unbinding path at different unbinding times transiting under the 64-72 loop. This was unexpected, given the rather linear shape of the GK allosteric pocket and the larger size of **4** and **5** (Fig. 3.8). Analyzing the unbinding processes in detail, we observed that **4** and **5** began dissociating when the rigid scaffold drifted towards the adjacent loop connecting β 10- β 11 sheets (241-250) or α 13 helix (444-455). Consequently, the residues located at the C-terminal and Tyr61, Arg63 sidechains were induced to shift from the initial position promoting the ligand solvation. Conformational rearrangements of the protein regions surrounding the dissociating ligand, including 64-72 loop, allowed the compound to completely unbind by transiting under the allosteric loop. We noticed that rotations of the flexible phenyl ring substituents were involved in driving 64-72 loop conformational

transitions and $\alpha 13$ helix displacement. Indeed, longer unbinding times were observed when **4** and **5** were sterically hindered into the allosteric binding pocket due to slow 64-72 loop structural rearrangements.

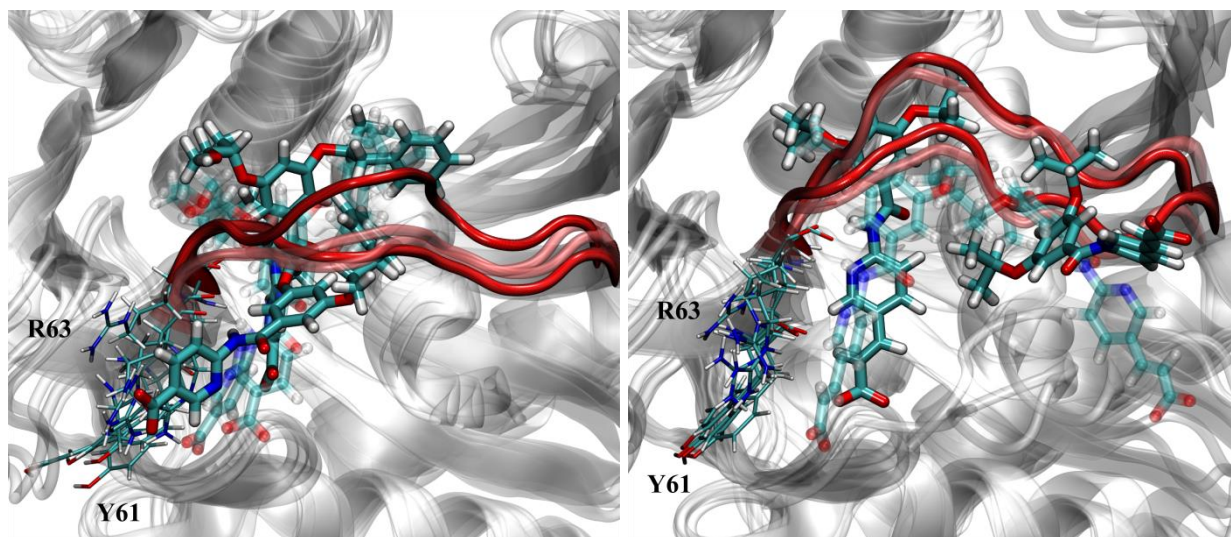


Figure 3.8. Representative unbinding paths explored by the **4** (Right) and **5** (Left). The 64-72 loop is red colored. The protein residues involved in the dissociation mechanisms are represented in stick.

In contrast to other approaches, such as sMD, the eLABMD approach does not require restraints on protein structure. Moreover, by minimizing the system perturbation, ABMD maintains the adiabatic character of the transformation. We exploited these peculiarities to carry out an in-depth mechanistic analysis of our unbinding eLABMD trajectories. Interestingly, when comparing the dissociation paths explored by structurally different GKAs, we identified conformational rearrangements of the CL1 portion facing the allosteric binding site (namely the 64-72 loop) as one of the rate-limiting steps of the GKA unbinding kinetics, influencing both unbinding times and paths. Two significant structural motions were observed, mainly during the dissociation of small T-shaped ligands. In one case, the 64-72 loop adopted a conformation stabilized by the Arg63 sidechain interacting with Asp158 and Asp160 included in the small lobe. Alternatively, the 64-72 loop moved towards $\alpha 5$ helix (205-215) allowing the establishment of the direct Glu67-His218 interaction. Large conformational changes accompanying the dissociation of the T-shaped ligands could be related to a retrograde induced-fit mechanism,^{12, 192} during which the protein operated the reverse sequence of structural rearrangements, leading to association by an induced-fit binding mechanism (e.g. occluding the binding site from bulk solvent). Thus, the experimental and computational slower off-rates associated with small branched GKAs might be attributed to the presence of electron donor groups in strategic points of the binding pocket, which confer stability to complexes and thus delay CL1 structural rearrangements. In contrast, GKAs displaying more pronounced linear shape and faster off-rates require moderate CL1 conformational rearrangements to dissociate, exploiting the rather linear shape of the allosteric pocket. The presence of highly flexible substituents instead of rigid aromatic moieties in larger T-shaped compounds increased the steric hindrance of ligands. That meant that structural rearrangements of the protein regions surrounding the binding site were necessary for unbinding. Nevertheless, the increased volume of ligand substituents had a limited effect on unbinding times. From our analysis, small T-shaped

geometry emerged as the most suitable shape for designing effective GK activators in terms of unbinding kinetics. One key factor in improving GKA residence times was the establishment of specific protein-ligand interactions, which constrained the conformation adopted by CL1 when the enzyme is in its active state. This confirmed the previous hypothesis.

3.4.2. A prospective application of elABMD protocol: the GSK-3 β case

3.4.2.1. Prospective predictions

Having assessed the ability of the elABMD-based protocol to retrospectively prioritize series of non-congeneric compounds coherently with experimental k_{off} , we challenged our approach in a prediction task.

We considered the recent work of Berg,¹⁶⁸ who designed pyrazine derivatives to cover a wide chemical space, making it possible to investigate the structural basis of the potency against GSK-3 β (i.e. K_i) and the selectivity versus cyclin-dependent kinase 2 (CDK2) (i.e. a protein kinase characterized by close homology to GSK-3 β). We selected and synthesized a series of six strictly congeneric ATP-competitive GSK-3 β inhibitors (see Table 3.2 for chemical structures), including three co-crystallized complexes (PDB IDs: 4ACG, 4ACD, 4ACH; all resolutions 2.6 Å), whose ligands (**1**, **2**, and **3**, respectively) showed significantly different experimental potency, K_i (0.22 nM, 4.9 nM, and 22 nM, respectively). Then, we chose three additional compounds from the series explored by Berg, looking at structural similarity and potency. Referring to the numbering of Table 3.2, we identified **4** (K_i = 0.67 nM), **5** (K_i = 12 nM), and **6** (K_i = 90 nM) as valid candidates to be included in our series, maintaining high levels of structural similarity enlarging the potency range of interest. We purposely co-crystallized these ligands to run our simulations (PDB IDs: **4** – 6HK4, **5** – 6HK3, **6** – 6HK7), as highly reliable binding configurations are mandatory to run unbinding simulations with MD-related approaches. Then, all the inhibitors were investigated by means of surface plasmon resonance (SPR) to experimentally determine the k_{off} values and validate our kinetic predictions. The presence of the morpholine group in **4** determines the loss of the positive charge, and a potency increase by a factor of 7 in comparison to its methylpiperazine analogue, **2**. The hydrophobic tail of **3** was shortened by one carbon atom in **6** causing potency to drop by a factor of 4. Understanding how slight structural differences influenced unbinding kinetics allowed us to thoroughly investigate the mechanisms that caused pairs of compounds with high structural similarity to have unexpectedly different off-rates, namely “kinetic cliffs”.

To prospectively prioritize the highly congeneric GSK-3 β series **1-6** on residence time, a high level of accuracy was required. To this purpose, we decreased the force constant with respect to GK to minimize the system perturbation increasing the accuracy of simulations. An extensive statistics of 20 independent replicas of 30 ns each was collected for each GSK-3 β complex setting the force constant equal to 4.0E-17 [kJ mol⁻¹]⁻³ to improve the robustness of predictions. In Table 3.7 and Figure 3.9, we report the unbinding time-based ranking correlations, which resulted from bootstrapped estimations of elABMD mean unbinding times, in

comparison to our experimental kinetic data collected by SPR (Table 3.6). Computational estimates were in good agreement with experimental residence time data. Three groups of GSK-3 β inhibitors with increasing predicted unbinding times were defined. The first group featured the most potent inhibitors of the series (**1**, **2**, **4**), the second group featured **3** and **5**, and the third group featured the weakest compound **6**.

In the Appendix, we report the rankings obtained by a random-selected statistics of 10 replicas, further demonstrating the robustness and inexpensiveness of the method. An additional test case (ligand) is also reported.

Table 3.6. Experimental kinetic data for each compound of GSK-3 β series^a

Cpd	K_D st-st	K_D kin	k_{on}	k_{off}	$t_{r,exp}$
1	2.79 ± 0.04	1.64 ± 0.01	$1.001 \pm 0.003E+06$	$1.64 \pm 0.01E-03$	609.76
2	8.4 ± 0.1	4.29 ± 0.02	$8.97 \pm 0.03E+05$	$3.85 \pm 0.02E-03$	259.74
3	17.75 ± 0.09	10.90 ± 0.02	$8.19 \pm 0.01E+05$	$8.925 \pm 0.006E-03$	112.04
4	2.81 ± 0.04	1.53 ± 0.01	$1.667 \pm 0.005E+06$	$2.55 \pm 0.02E-03$	392.16
5	31.4 ± 0.6	18.93 ± 0.05	$2.441 \pm 0.006E+05$	$4.565 \pm 0.005E-03$	219.06
6	45.9 ± 0.3	42.3 ± 0.2	$1.286 \pm 0.005E+05$	$5.4400 \pm 0.0001E-02$	18.38

^a K_D values (both K_D steady-state and K_D kinetics) are expressed in nM, k_{on} in $s^{-1} M^{-1}$, k_{off} in s^{-1} , experimental residence times, $t_{r,exp} = 1/k_{off}$, in s. The experimental errors refer to the quality of the model fit.

Table 3.7. Mean estimation of elABMD unbinding times ($t_{r,mean}$) and experimental kinetic data for each compound of GSK-3 β series^a

Cpd	<u>elABMD</u>		<u>Experimental</u>	
	$t_{r,mean}$	Rank $t_{r,mean}$	$t_{r,exp}$	Rank $t_{r,exp}$
1	28.64 ± 0.60	1	609.76	1
2	26.53 ± 1.56	3	259.74	3
3	25.72 ± 1.40	4	112.04	5
4	28.45 ± 0.77	2	392.16	2
5	24.94 ± 1.70	5	219.06	4
6	18.75 ± 2.02	6	18.38	6

^a eLABMD unbinding times ($t_{r,mean}$) are expressed in ns and reported with an estimation of the error computed with a bootstrapped procedure. Experimental residence time are expressed in s. Spearman coefficient for mean eLABMD unbinding time-based ranking correlations with respect to $t_{r,exp}$ is equal to 0.94.

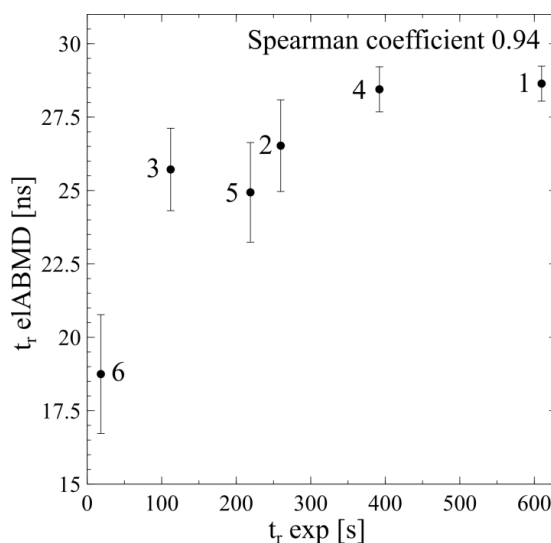


Figure 3.9. Experimental residence time ($t_{r,exp}$) *versus* computational unbinding time. The ranking obtained with bootstrapped estimations of mean eLABMD unbinding times against $t_{r,exp}$ is reported.

3.4.2.2. Explanation of protein-ligand unbinding paths

Looking at the collected trajectories, we noticed that all GSK-3 β inhibitors included in the chemical series explored one consistent unbinding path directed towards the small N-lobe, possibly preceded by slight rotations of the scaffold inside the ATP-binding pocket. In Figure 3.10, the minimum distances registered for **2** during a representative unbinding trajectory are reported against the protein residues. It suggests that the ligand spent most of the simulation time into the binding pocket and it dissociated without being significantly involved in side interactions with protein residues around the binding site. All binding poses were initially destabilized by the upward movements of the solvent-exposed methylpiperazine group, or by the fluctuations of substituents to the amide group leading to structural rearrangements of the glycine-rich loop, or a combination of both. Due to simulations parameters, dissociating ligands could establish side interactions with protein residues surrounding the binding site.

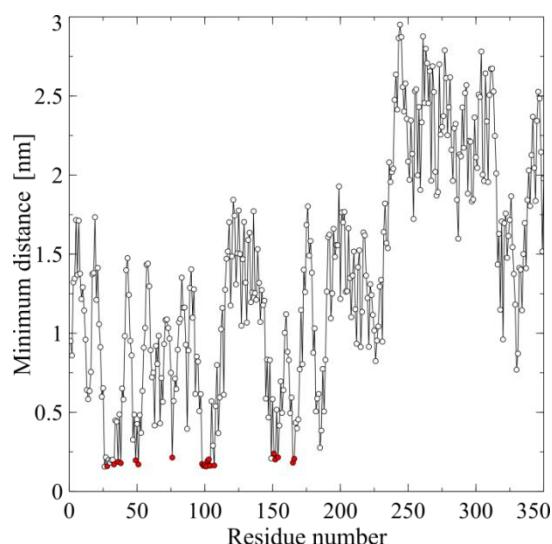


Figure 3.10. Minimum distances explored by **2** during a representative unbinding trajectory against protein residues. The numbering of residues is reported on the abscissa. The minimum distance between protein and ligand is reported in nm on the ordinate. The selection of protein residues into the binding site are highlighted in red. See text for details.

Despite the selected GSK-3 β inhibitors belong to the same scaffold, they bear different functional groups. To understand how these differences affect the unbinding kinetics, we made an in-depth analysis of the dissociation mechanisms explored by couple of analogues (i.e. structure-kinetic relationship, SKR).

We evaluated the influence of the positively charged methylpiperazine group on dissociation dynamics by comparing the unbinding paths explored by **2** and its morpholine analogue **4**. During the dynamics, the morpholine oxygen atom of **4** turned to the Lys60 sidechain, optimizing the twist of the adjacent phenyl ring and the resulting interactions with Ile62 and Arg141. Hence, wide fluctuations of the morpholine group and unfavorable motions of the scaffold inside the binding pocket were prevented blocking the access of water molecules into the binding site (Fig. 3.11, Left). Therefore, the complete solvation of **4** required the rupture of interactions among pyridine ring, Lys85, and Phe67, in addition to the clockwise rotation of the scaffold, breaking the key hydrogen bonds with the hinge residues. The presence of the polar methylpiperazine group limited the stabilization of the positively charged ring to water-mediated interactions involving Glu137 and Thr138 preventing the establishment of a stable interaction with Lys60 sidechain. The resulting fluctuations of the solvent-exposed methylpiperazine group led to positioning rearrangements of **2** promoting the access of water molecules into the ATP-binding site and faster dissociation (Fig. 3.11, Right).

Figure 3.12 reports the distance between the nitrogen atom of Lys60 sidechain and the nitrogen atom of the methylpiperazine ring of **2** and the oxygen atom of the morpholine ring of **4**. The lower distance between Lys60 and **4** suggests an increased resistance and a stabilized bound state.

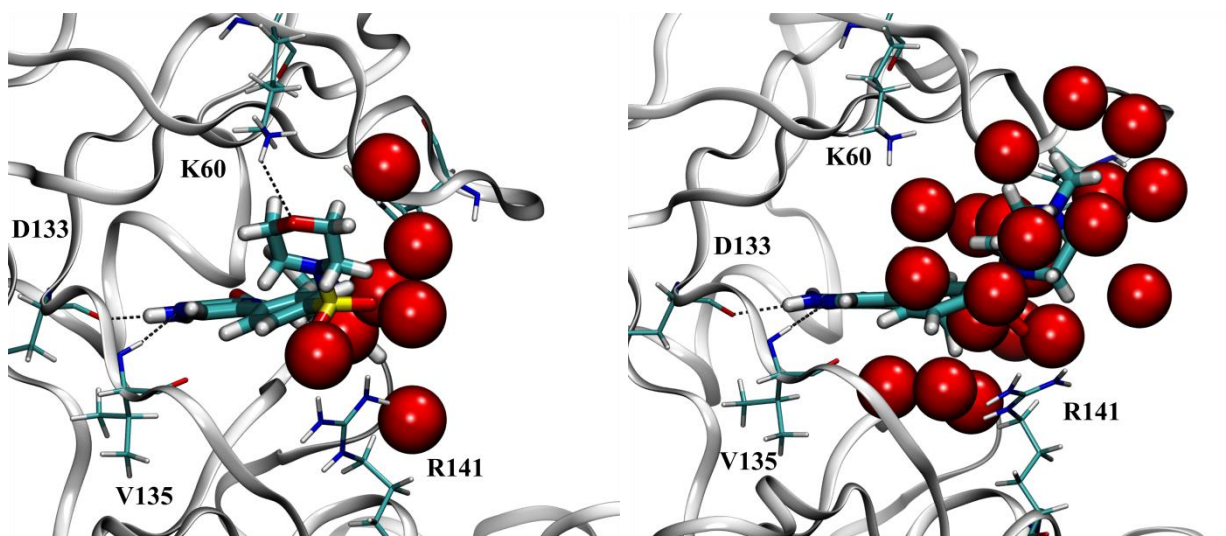


Figure 3.11. Illustrative representation of distribution of water molecules around **4** (Left) and **2** (Right). The hydrogen bond established between the oxygen atom of morpholine ring of **4** and Lys60 sidechain prevented the access of water molecules inside the ATP-binding pocket.

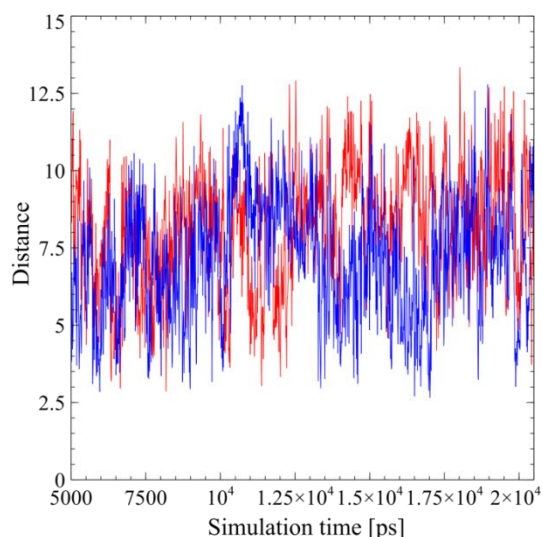


Figure 3.12. Fluctuations of distance between the nitrogen atom of Lys60 sidechain and the nitrogen atom of the methylpiperazine ring of **2** (red) and the oxygen atom of the morpholine ring of **4** (blue) were plotted against the simulation time. Two representative unbinding trajectories were considered in the reported time range corresponding to ligands' bound states. The distance is reported in Angstrom.

In comprehensively evaluating all unbinding paths leading to complete ligand solvation, we noted that the computed unbinding time increased when the motions of functionalities connected to the amide group were limited by the establishment of hydrogen bonds with the conserved Lys85 or by intra-molecular interaction with the adjacent amide nitrogen atom. Nitrogen atoms of pyridine ring characterizing **1**, **2**, and, **4** directly interacted with Lys85 sidechain. This happened for the terminal methoxy functionality of **6** to a lesser extent (Fig. 3.13, Left). The bound states of **3** and **5** were stabilized by intra-molecular interactions, engaging the methoxy functionality and the adjacent amide nitrogen atom (Fig. 3.13, Right). In both cases, the resulting dissociation mechanisms were dependent on structural rearrangements of glycine-rich loop opening the flexible binding site.

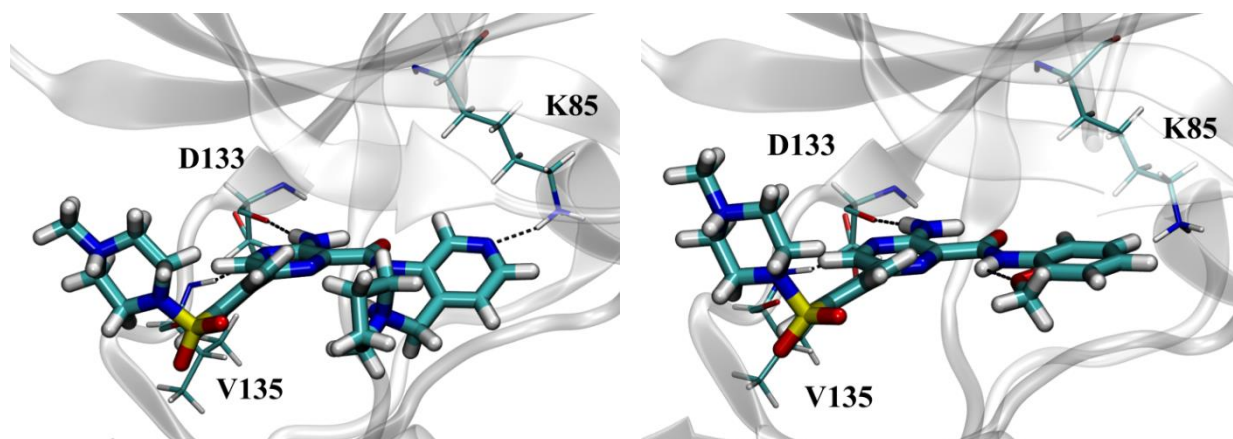


Figure 3.13. Different stabilizations of the substituent to the amide group. **1** (Left) establishes a hydrogen bond with Lys85 sidechain. **5** (Right) keeps the planarity of the methoxy-phenyl ring through the establishment of an intra-molecular hydrogen bond.

We compared the dissociation mechanisms explored for **2** and **5** in order to qualitatively evaluate how the unbinding kinetics was influenced by the direct hydrogen bond with Lys85 sidechain, and by the intra-molecular interaction with the amide nitrogen. The bound state stability was affected by the additional internal degree of freedom due to the rotatable bond of the methoxy substituent in ortho position. This facilitated the rupture of the internal hydrogen bonds when hydrophobic stacking interactions with Phe67 were weakened. In contrast, **2** maintained the hydrogen bond with Lys85, despite slight rotations of the pyridine ring distorting the planar conformation to the amide group. Therefore, structural rearrangements of the scaffold inside the binding site were limited by hydrogen bonds with conserved Lys85, increasing the unbinding time.

1 and **3** had in common the folded conformation of substituents to the amide group, which partially filled the ribose binding site, burying the hydrophilic moiety towards the inner part of the binding pocket while exposing the hydrophobic one to the solvent. Interestingly, this site was occupied by a water molecule in the crystal structure of GSK-3 β in complex with **2**. This water molecule was absent in experimental complexes of **1** and **3**, suggesting a gain in potency due to the displacement of a high-energy water molecule. During the dynamics, pyrazine ring of compound **1** stably interacted with Lys85, whereas the pyrrolidine ring was free to move, establishing water-mediated interactions with C-lobe residues, such as Gln185 and Asn186, or with glycine-rich loop when pointed the N-lobe. Alternative orientations of the pyrrolidine group transiently left the partially filled ribose binding site, attracting water molecules towards the position occupied by the crystallographic water molecule. Ligand solvation was initiated by upward motions of the **1** scaffold, mainly due to methylpiperazine group fluctuations combined with structural rearrangements of glycine-rich loop. In contrast, the weakly stabilized folded conformation of methoxypropyl chain of **3** rapidly adopted an extended conformation. This results in the solvent-exposure of the hydrophilic methoxy group, which allowed water molecules to partially fill the ribose binding site. This cancelled the gain in potency acquired by the displacement of the high-energy water molecule necessary for binding.

To test the predictive ability of eLABMD protocol and its applicability to highly congeneric chemical series, we simulated and analyzed the dissociation processes of a selection of GSK-3 β ATP-competitive inhibitors with subtle structural modifications and different experimental potencies. After establishing that the predicted unbinding-time-based ranking correlations and experimental data were in agreement, we analyzed how subtle structural modifications affected the unbinding mechanisms and ligand solvation. From the observations of all trajectories, we identified one homogeneous unbinding path. We then focused on substituents that differentiated the chemical series. We thus deepened the structural basis of different experimental potency values. We observed that hydrophilic substituents in specific positions increased the duration of the protein-ligand complex, reducing structural rearrangements of protein residues around the binding site and limiting the degree of accessibility of water molecules.

3.5. Discussion and conclusions

In the present work, we used a computational protocol to simulate protein-ligand dissociation events. This protocol combines adiabatic-bias molecular dynamics with an electrostatics-driven collective variable, dubbed eLABMD. The accurate eLABMD unbinding simulations allowed us to evaluate dissociation processes where the rate-limiting step was related to conformational rearrangements of protein domains (GK) or to solvation effects due to the determinants of particular ligands. The analysis thus identified structural modifications that would be suitable for improving the duration of the protein-ligand complex. In summary, the proposed methodology, together with a qualitative description of the unbinding process, can deliver quantitative information in a very modest computational time. It should therefore be useful to the scientific community.

Adiabatic bias molecular dynamics (ABMD) was first applied by Bortolato et al.¹⁵⁸ to simulate ligands unbinding, combined with well-tempered metadynamics¹⁶⁰ to estimate the energy barrier required to start the dissociation. In metadynamics, generic path collective variables¹⁹³ are considered to define the root-mean-square deviation (RMSD) position on the path (s) and the RMSD distance from the path (z). By definition, metadynamics gradually enhances the probability of visiting different states. Thanks to ABMD, only the states compatible with ligand unbinding events were explored, speeding up the simulations. The results provided a physics-based, fully flexible, and pose-dependent ligand scoring function, namely residence time score (RTscore), which quantified the maximum energy bias (i.e. transition state energy) required to induce the transition of the ligand from an energy basin to another. Under the assumption that the first steps of the unbinding events would represent the principal kinetic bottleneck, the resulting RTscore values were assumed to be related to residence times, and then used to prioritize the chemical series.

In contrast to Bortolato, we did not use metadynamics and we did directly leverage the information due to the unbinding time. Indeed even if the time is a biased quantity, the general strategy of observing rates or time is much more direct. We expected it to be more effective than reconstructing the free energy, which is a much more complex and exquisitely inverse problem. We chose a collective variable able to monitor the electrostatic potential between interacting entities, instead of the RMSD of the bound/unbound states of the ligand. The idea was to improve the description of the natural forces driving dissociation mechanisms, thus making our approach more physics based and also avoiding the need to define an arbitrary unbound pose. Additionally, our focus was on extrapolating mechanistic and path information on unbinding events. Therefore, fully flexible unbinding trajectories were analyzed from a mechanistic point of view, in addition to estimating the dissociation time.

To further validate the reliability of our unbinding kinetics predictions, we are working on the correction of the biased unbinding times in order to obtain the absolute, physical residence times. The idea is to quantify the work due to bias that has been applied to facilitate the ligand unbinding, with respect to the force constant, K . The rigorous correction can be established through the Jarzynski equality.¹⁹⁴⁻¹⁹⁵ Note that the amplitude of the bias changes among the ligands included in the series and also among the several replicas performed for the same compound, making this evaluation not straightforward, but essential.

To complete the characterization of the unbinding process, free energy considerations are needed. To this aim, we are developing a computational approach to estimate the potential of mean force of the unbinding process in a semi-automated way. The protocol to obtain the PMF has two main phases. First, a sequence of waypoints along the path is placed by a machine learning algorithm producing a smooth string from an initial noisy unbinding trajectory. Second, an optimization protocol based on a series of subsequent steered MDs has been developed to uniform the spacing between consecutive frames in the path, a necessary condition to apply the path collective variables (PCVs).¹⁹³ The free energy surface (FES) explored by the ligand during the dissociation event is then reconstructed by *w*-META-D. By fixing the value of the Z -variable (i.e. the distance from the path), it is possible to define the unbinding free energy profile as a potential of mean force (PMF). By defining the PMF, this approach would allow the identification of the binding intermediates and give insights into transient drug-target molecular interactions, in turn driving the rational optimization of lead compounds on binding/unbinding kinetics.

4. A computational approach to estimate absolute free energies and hydration free energies in atomistic simulations

4.1. Aim of the project

Thermodynamics and kinetics of every biomolecular system ultimately depend on the potential energy landscape through an averaging process that relies on the statistical mechanics ensemble used in their definition. The former provides the energetic binding force and the latter describes the rates of transition between energy basins. Even though thermodynamics and kinetics are usually described separately, they both relate to free energy. Free energy considerations are of particular interest for the interpretation of the equilibrium properties and of the kinetic transformations in terms of microscopic interactions finding application in various research fields, including drug discovery.¹⁹⁶

4.2. Flow diagram of our free energy computation

In this chapter, we aim at computing absolute free energies of water-solute systems, as well as the hydration free energies of the same solute molecules.

Hydration free energy is estimated as the free energy difference of the water-solute sample, pure water, and one solute molecule, all computed at normal conditions. Therefore, we dealt with fluid water samples and with pure solute systems in the liquid, solid, and gas/vapor phases.

Our computational scheme relies on the general strategy applied to compute the absolute free energy of each sample (i.e. water-solute, pure water, pure solute) including the quasi-harmonic term, $F_{harm}(V, T)$, arising from the vibrational normal modes of the system, the ideal contribution, F_{id} , and the an-harmonic term, ΔF , computed by thermodynamic integration (TI), which recovers the an-harmonicity of the real system.

$F = F_{harm}(V, T) + F_{id} + \Delta F$	(4.01)
--	--------

Our computational scheme requires the preparation of a suitable sample, which is subsequently equilibrated at room temperature and low pressure by molecular dynamics with the Langevin thermostat. Equilibration lasts ~100 ps. Each sample is briefly annealed from 300 K down to 150 K during 20 ps, and then quenched to the nearest local energy minimum. The short annealing is introduced to enhance the correlation among the local minima found for the different systems, while driving the system towards an amorphous configuration. The sudden energy minimization is carried out by quenched MD that sets velocities to zero whenever the system tends to move uphill in energy ($\sum_i \mathbf{v}_i \cdot \mathbf{F}_i < 0$) and confines the system to the starting energy basin.

Samples obtained in this way are used to compute and diagonalize the dynamical matrix, giving eigenvalues and eigenvectors of vibrational normal modes. Eigenvalues, in particular, give access to the harmonic free energy of the system, $F_{\text{harm}}(V, T)$. This step takes a few hours on a single-core CPU, but it might grow to a sizeable computation for systems sizes exceeding the few thousand atoms, unless steps are taken to limit its cost (Sec. 4.5). For extended systems, this basic step of harmonic dynamics and thermodynamics is repeated ~ 10 times for a series of volumes differing by $\sim 1\%$ from each other to find the volume of minimum free energy at 300 K, thus completing the quasi-harmonic (QH) step for that sample. This stage is eased by the fit of the volume-dependent free energies by an analytic equation of state (i.e. Birch-Murnaghan equation of state; Sec. 4.3.5). This procedure to estimate the QH free energy is applicable to fluid and solid samples, but not to pure solutes that are in gas phase at normal conditions. In those cases, the free energy is estimated from grand-canonical Monte Carlo (GC-MC) simulations (Sec. 2.1.4 and 2.5).

The eigenvalues and eigenvectors at the volume of minimum free energy and at $T = 300\text{ K}$ are fed to the TI step. At this stage, ~ 16 values of the interpolating parameter $0 \leq \lambda \leq 1$ are introduced to set-up the Hamiltonian $H(\lambda)$, and compute the corresponding $\langle \partial H(\lambda) / \partial \lambda \rangle_\lambda$, as discussed in Section 4.3.6. Tabulated values spanning the $\lambda \in [0: 1[$ interval and excluding $\lambda = 1$, are integrated in λ by the trapezoidal rule. The placement of the points is discussed in Sections 4.3.6 and 4.4.2.

To compute the absolute free energy, the entire $[0: 1]$ range has to be covered by the integration. However, the integrand is singular at $\lambda = 1$, hence the trapezoidal integration sets the $\lambda = 1$ integrand at the value of the highest λ point that can be computed with an acceptable error bar. In practice, this corresponds to $\lambda = 0.995$, and, since the integral remains finite despite the divergence of the integrand, this provides a useful approximation to the sought for absolute free energy.

The computation of free energy differences, such as hydration free energies, benefits of the expected near cancellation of the integrand when taking the free energy difference between systems that are similar in size, such as solute in water and bulk water (Sec. 4.4.3).

Under the assumption of perfect cancellation at $\lambda \geq 0.95$ (i.e. at long range), the trapezoidal integration covers the $[0: 0.995]$ range only. Hence, the ΔTI contribution computed for each pair of samples, such as solute in water and bulk water, is displayed up to $\lambda = 0.995$ (Table 4.8a-c).

Further details on the way to account for the dissociation/protonation states of solutes, on approaches to prevent the $\lambda = 1$ divergence of the integrand, and on strategies to improve the scaling of the computation for large systems are discussed in the following sections.

4.3. Simulation setup

4.3.1. The models

The parameters of the SPC/Fw (flexible simple point charge)¹⁹⁷ water model were used to describe the potential energy surface of liquid bulk water. Note that a recent version of the Amber force field (i.e. ff15ipq) is primarily tailored for the rigid SPC/Eb¹⁹⁸ water model, justifying our choice.

In Table 4.1, the parameters of the SPC/Fw water model are reported.

Table 4.1. Parameters of the SPC/Fw water model^a

Model	$k_{stretching}$	r_{OH}	$k_{bending}$	$\cos \theta$	σ_{OO}	ϵ_{OO}	q_O	q_H
SPC/Fw	4431.53	1.012	376.13	-0.395	3.165	0.650	-0.82	0.41

^a Energies are quoted in kJ/mol, distances in Å, angles in rad. Charges are in units of |e|.

Globally neutral small molecules were modeled according to the GAFF (General Amber force field; Sec. 2.2.3.1)⁶³ force field. In particular, the small molecules were parametrized using the same GAFF parameters applied by Mobley et al. and reported in the FreeSolv database.¹⁹⁹

The functional form of the GAFF force field is:

$ \begin{aligned} U_{GAFF} = & \sum_{bonds} k_{str}(r - \bar{r})^2 + \sum_{angles} k_{\theta}(\theta - \bar{\theta})^2 + \sum_{dihedrals} \sum_{n=1} \frac{V_n}{2} (1 + \cos(n\phi - \gamma)) \\ & + \sum_{i < j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \end{aligned} $	(4.02)
--	--------

Improper torsions²⁰⁰ might be added to enforce the planarity of chemical groups.

4.3.2. Units

The system of units used in our computations is defined by the following choices:

Unit of length, L: the Angström = 10^{-8} cm

Unit of mass, M: the atomic mass unit = $1.660539 \cdot 10^{-24}$ g

Unit of energy, E: the kJ/mol = $1.640 \cdot 10^{-14}$ erg

With this choice, time is a derived unit:

	$[E] = \frac{[M][L^2]}{[t^2]} \rightarrow [t] = \sqrt{\frac{[M][L^2]}{[E]}} \cong 10^{-13} \text{ s}$	(4.03)
--	---	--------

4.3.3. System preparation

As a first step, a cubic box of 300 SPC/Fw (flexible simple point charge)¹⁹⁷ water molecules was built choosing the box size, \bar{V} , resulting in a density of 1.00 g/cm³. In the following Section 4.3.5, the procedure carried out to optimize the volume and density of the system applying the quasi-harmonic (QH) approximation is presented. During the production runs, Langevin thermostat²⁰¹ was applied to sample the NVT ensemble. The center of mass of the whole system is fixed at the origin, but the center of mass of every single molecule (solute included) is unconstrained.

We considered a short series of globally neutral small solutes listed in Table 4.2, whose hydration free energy (HFE) has been previously determined both by experiments and computations. The chemical series was enlarged including a drug molecule (ketoprofen). As experimental HFEs, we considered the values summarized in the work of Martins et al.²⁰² The experimental HFE of ketoprofen refers to the work of Geballe and coworkers.²⁰³ As theoretical benchmark, the FreeSolv database by Mobley et al.¹⁹⁹ was used to validate our predictions.

Table 4.2. Experimental (HFE_{exp}) and computational (HFE_{FreeSolv}) hydration free energies of our series of globally neutral small molecules^a

Solutes	HFE _{exp}	HFE _{FreeSolv} ¹⁹⁹
Methane	8.24	10.25
Isobutane	9.67	10.63
Nitromethane	-16.74	-8.70
Benzene	-3.64	-3.39
Propionic acid	-27.03	-38.03
Piperidine	-21.42	-16.19
Ketoprofen	-45.10	-72.13

^a The experimental HFEs²⁰² of propionic acid, piperidine, ketoprofen refer to their charged forms. The computational HFEs refer to the FreeSolv database. All energy terms are expressed in kJ/mol.

Periodic boundary conditions (PBC) were applied to minimize boundary effects by finite size, and to approximate the infinite system by the simulated one. Following standard protocols, short-range Lennard-Jones-like non-bonded interactions were cut at 3.00σ . Long-range Coulomb interactions were dealt with by the Ewald summation method.⁷³ Screened Coulomb interactions have been cut at the range of the O-O dispersion interactions.

Every pair potential is switched off smoothly at the end of their finite range. To this aim, the potential is multiplied by a cubic polynomial function within a spherical corona of width 0.4 \AA added to the interaction sphere. As a result of this procedure each potential vanishes with its first derivative within a finite range. Newton's equations of motion were integrated using the velocity Verlet algorithm⁷¹ fixing the time step at 1 fs. All calculations were carried out using programs written from scratch in FORTRAN. Covalent bonds involving hydrogen atoms were left unconstrained throughout the simulations. Each system was initially equilibrated for 700 ps at 300 K in the NVE ensemble.

To carry out the harmonic and quasi-harmonic steps, the system was driven to the nearest local minimum by quenched molecular dynamics. To this aim, one representative configuration of the system was shortly simulated (15 ps) annealing it from 300 K to 150 K before quenching, in order to improve the correlation among the local minima obtained for different samples. The quenched configuration was used as reference configuration defining positions of local equilibrium of all atoms $\{\mathbf{R}_i^0\}$. Then, the Hessian matrix was computed and diagonalized as described in the following Section 4.3.4. In the Results section, the suitability of an arbitrary local minimum conformer as reference configuration is discussed.

4.3.4. Computation and diagonalization of the Hessian matrix

By definition the Hessian is a $(3N \times 3N)$ matrix whose elements are the partial second derivative of the energy with respect to the coordinate of the $3N$ atoms i and j . As such, it can be interpreted in terms of harmonic springs connecting pairs of atoms as:

	$k = \frac{\partial^2 U}{\partial R_i^\alpha \partial R_j^\beta}$	(4.04)
--	---	--------

where U is the potential energy of the force field, i, j labelling atoms, and α, β being Cartesian components (x, y, and z).

In our approach, the Hessian matrix is computed by numerical differentiation fixing Δ equal to 10^{-5} \AA . For $i \neq j$ or $\alpha \neq \beta$:

$\frac{\partial^2 U}{\partial R_i^\alpha \partial R_j^\beta} = \frac{U(R_i^\alpha + \Delta; R_j^\beta + \Delta) - U(R_i^\alpha + \Delta; R_j^\beta - \Delta) - U(R_i^\alpha - \Delta; R_j^\beta + \Delta) + U(R_i^\alpha - \Delta; R_j^\beta - \Delta)}{4\Delta^2}$	(4.05)
---	--------

Diagonal elements ($i = j$ and $\alpha = \beta$) are computed as:

$\frac{\partial^2 U}{\partial (R_i^\alpha)^2} = \frac{U(R_i^\alpha + \Delta) + U(R_i^\alpha - \Delta) - 2U(R_i^\alpha)}{\Delta^2}$	(4.06)
--	--------

This is the simplest finite difference approximation to second order partial derivatives. More sophisticated approximations are available and implemented in quantum-mechanical (QM) chemistry packages to compute vibrational modes.

Around each minimum, the harmonic Hamiltonian, $H_{\text{harm}}(\{P_i\}; \{R_i\})$ is defined as:

$H_{\text{harm}}(\{P_i\}; \{R_i\}) - E_0 = \sum_i \frac{P_i^2}{2M_i} + \frac{1}{2} \sum_{\alpha, \beta} \sum_{i, j} \{R_i - \bar{R}_i\}_\alpha \left(\frac{\partial^2 U}{\partial R_i^\alpha \partial R_j^\beta} \right) \{R_j - \bar{R}_j\}_\beta$	(4.07)
--	--------

After transforming the Hamiltonian into the dynamical matrix, the harmonic frequencies, ω_i , are provided by diagonalization, and the vibrational density of state (vDOS) of the system is computed. As a technical detail, we point out that for periodic systems the eigenstates obtained in this way represent the Γ point only vibrational properties of the system. The fairly large size of the supercell ensures that the approximation is accurate.

Note that the cost of computing and diagonalizing the Hessian matrix represents a limitation in the application of the approach to biological systems of pharmacological interest including thousands of interacting atoms. In Section 4.5, approaches to optimize the computational machinery are introduced.

4.3.5. Volume optimization by quasi-harmonic (QH) approximation

The volume of the system was optimized applying the quasi-harmonic (QH) approximation through small variations (1%) of the system volume. In each volume, particles were equilibrated and quenched and the corresponding free energy was computed. Each quenched configuration was then used as reference conformer, from which the positions of local equilibrium were extracted. Therefore, for each volume of the system, the corresponding Hessian matrix was computed and a set of volume-dependent harmonic frequencies were provided allowing the minimization of the harmonic free energy for each temperature with respect to the volume (Fig. 4.1). At this minimum, the pressure $P = -(\partial F / \partial V)_{T, N}$ is zero by definition.

The classical harmonic free energy as a function of the volume-dependent frequencies is expressed as:

$F_{harm}(V, T) = -k_B T \ln Q = -k_B T \sum_{i=1}^N \ln \left\{ \frac{k_B T}{\hbar \omega_i(V)} \right\}$	(4.08)
--	--------

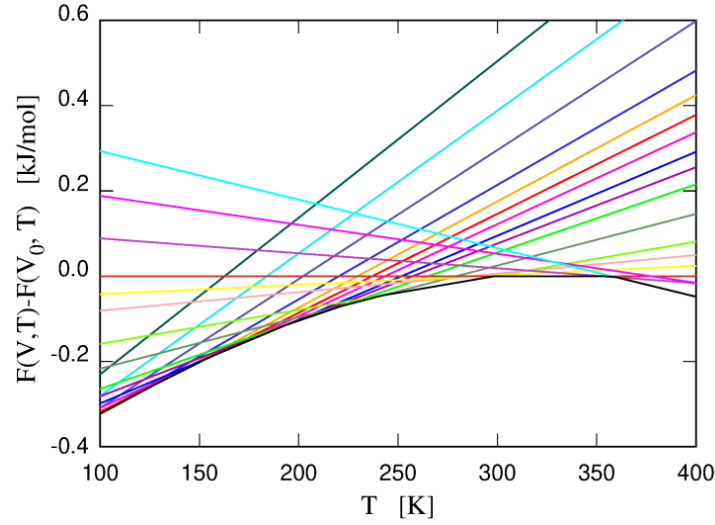


Figure 4.1. Illustration of the computation of the free energy per molecule in the quasi-harmonic approximation for the system of 300 SPC/Fw water molecules. Each line in color represents the free energy as a function of the temperature for a given volume. The system free energy is the lower convex envelope of the curves, drawn in black. This free energy corresponds to a range of volumes.

The hydrogen bond is directional and mechanically “fragile”, and then the potential energy surface is rugged, marked by a multitude of nearly equivalent H-bonding configurations in water. A fit procedure is introduced to remove the dependence of the free energy surface on these details. To this aim, the QH free energy, $F_{harm}(V, T)$, at 300 K for each volume was interpolated by the Birch-Murnaghan isothermal equation of state.²⁰⁴

$P(V) = \frac{3B_0}{2} \left[\left(\frac{V_0}{V} \right)^{7/3} - \left(\frac{V_0}{V} \right)^{5/3} \right] \left\{ 1 + \frac{3}{4} (B' - 4) \left[\left(\frac{V_0}{V} \right)^{2/3} - 1 \right] \right\}$	(4.09)
--	--------

where P is the pressure, V is the volume, V_0 is the equilibrium volume at $P = 0$, B_0 is the bulk modulus at $P = 0$, and B' is the pressure derivative of the bulk modulus.

The pressure as a function of the free energy, F , is defined as:

$P(V) = - \left(\frac{\partial F}{\partial V} \right)_{N, T}$	(4.10)
--	--------

Equation 4.11 defines the bulk modulus, $B(V)$, which dimensionally is a pressure (GPa, in our computations).

	$B(V) = -V \left(\frac{\partial P}{\partial V} \right)_{N,T}$	(4.11)
--	--	--------

Thus, the expression for the free energy whose derivative is the Birch-Murnaghan equation of state is:

	$F(V) = E_0 + \frac{9V_0B_0}{16} \left\{ \left[\left(\frac{V_0}{V} \right)^{2/3} - 1 \right]^3 B' + \left[\left(\frac{V_0}{V} \right)^{2/3} - 1 \right]^2 \left[6 - 4 \left(\frac{V_0}{V} \right)^{2/3} \right] \right\}$	(4.12)
--	---	--------

This last expression was used to fit the free energy at 300 K obtained from the QH approximation, $F_{harm}(V)$, by a non-linear fit routine.

In practice, the QH equation of state has been determined interpolating the volume-dependent harmonic free energies covering a range of $0.94\bar{V} \leq V \leq 1.07\bar{V}$, where \bar{V} is the volume resulting in a density of 1.00 g/cm³.

The calculations were performed both in quantum mechanics and in the classical limit. Because of the zero point energy, the equilibrium volume, V_0 , obtained from the QM calculation is ~3% larger than its classical counterpart (Table 4.3, Fig. 4.2). Notice that a 3% volume difference corresponds to a lattice constant difference of just 1%, which is comparable to the uncertainty in the computational and experimental results for soft matter systems.

Table 4.3. Equilibrium volume, QH free energy, and bulk modulus resulting from the fitting of the QH free energy by the Birch-Murnaghan equation of state of the bulk water system including 300 SPC/Fw molecules^a

SPC/Fw water model			
Model	V_0	$F_{harm}(V_0)$	B
Classical	30.823	-32.142	3.453
Quantum	31.715	-62.869	3.456

^a The equilibrium volume, V_0 , and the corresponding quasi-harmonic free energy at 300 K, $F_{harm}(V_0)$, are expressed in Å³ and kJ/mol, respectively. The bulk modulus, B , is expressed in GPa. Quantum estimate of E_0 does not include zero point energy. Energy values refer to the single water molecule.

The density corresponding to V_0 is 0.97 g/cm³ in classical mechanics and 0.94 g/cm³ in quantum mechanics.

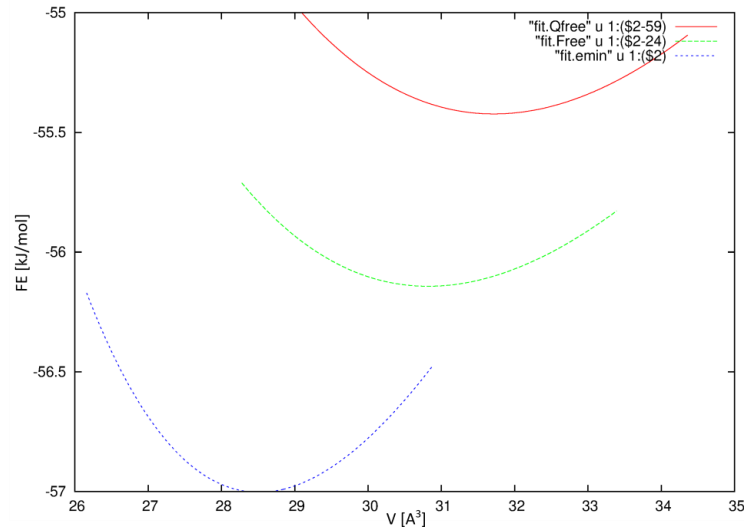


Figure 4.2. Potential energy and free energy of water as a function of the volume. The curves were shifted on the y-axis to help their comparison. The classical harmonic free energy at $T=0$ K is blue colored, the classical and quantum free energies at $T=300$ K are reported in green and red, respectively. The shift on the x-axis shows the thermal expansion of the system moving from $T=0$ K to $T=300$ K. The shift is more pronounced in the quantum free energy than in the classical one.

The free energy data, $F_{\text{harm}}(V, 300 \text{ K})$, from the QH approximation are scattered around the minimum volume, V_0 , especially on the high volume side, because of peculiarities of the network structure of water. By interpolating the data with Equation 4.12, these fluctuations were removed and a robust estimate of the equilibrium volume and QH free energy were provided (Fig. 4.3).

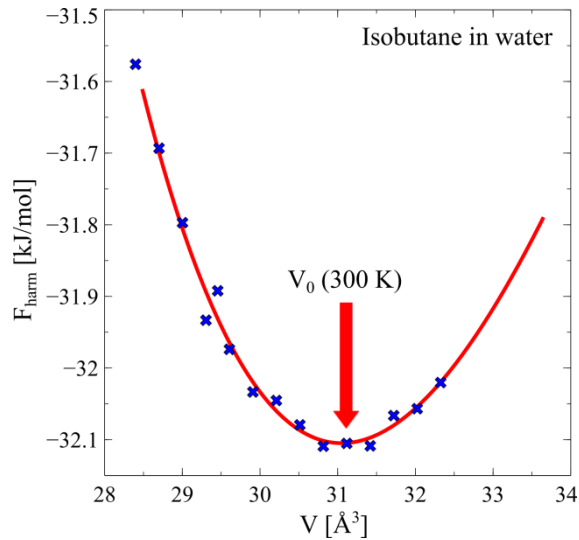


Figure 4.3. Equation of state of QH free energy at 300 K computed for the system including one molecule of isobutane solvated in 300 water molecules. The fit by the Birch-Murnaghan equation of state is red colored. The minimum represents the free energy estimate and the equilibrium volume in the quasi-harmonic approximation.

The equilibrium volume in the QH approximation as a function of the temperature is reported in Figure 4.4. Note that by minimizing the equilibrium volume identified by the QH approximation, the minimum volume of water at $T = 4$ °C is not reproduced being an intrinsically an-harmonic effect.

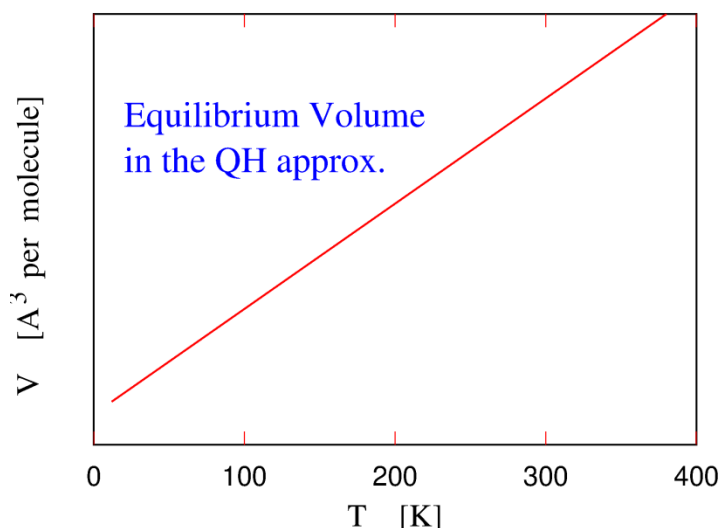


Figure 4.4. Equilibrium volume in the QH approximation as a function of the temperature. Classical mechanics computation.

A detailed analysis of the QH results on ketoprofen is reported in Section 4.4.3.1.

In the Appendix 7.7, the QH equilibrium volume in the classical approximation is validated on three representative systems (i.e. bulk water, nitromethane, benzene) by comparing the QH result with the equilibrium volume obtained by a single run at 300 K in the NPT ensemble.

4.3.6. Thermodynamic perturbation

The an-harmonicity characterizing fully interacting systems was recovered by thermodynamic integration⁸⁷ (TI) allowing the computation of the free energy difference between the reference (i.e. harmonic) and real systems. Since the kinetic part is always the same, the interpolation affects the potential energy only. The potential energy as a function of the perturbation parameter, λ , was defined as linear interpolation between two states setting $\lambda = 1$ for the description of the fully interacting system (i.e. force field, U_{ff}) and $\lambda = 0$ for the harmonic one (U_{harm}).

	$\Delta H(\lambda) = U(\lambda) = \lambda U_{ff} + (1 - \lambda) U_{harm}$	(4.13)
--	--	--------

With this choice:

	$\frac{\partial \Delta H(\lambda)}{\partial \lambda} = H_{ff} - H_{harm}$	(4.14)
--	---	--------

We compared two different implementations.

In the former, we first diagonalized the dynamical matrix and then we used the elongation dx_i along each eigenstate i as coordinate. Thus, the harmonic potential energy was defined as:

	$U_{harm} = E_0 + \frac{1}{2} \sum_{i=4}^{3N} \omega_i^2 dx_i^2$	(4.15)
--	--	--------

In the latter, we kept the original Cartesian atomic coordinates, and we computed the harmonic potential energy from the Hessian as:

	$U_{harm} = E_0 + \frac{1}{2} \sum_{\alpha, \beta} \sum_{i, j} (R_i - \bar{R}_i)_\alpha \frac{\partial^2 U}{\partial R_i^\alpha \partial R_j^\beta} (R_j - \bar{R}_j)_\beta$	(4.16)
--	--	--------

The implementation based on the elongation (Eq. 4.15) makes it easier to introduce variations from the harmonic Hamiltonian. The other implementation (Eq. 4.16) is suitable to enforce localization in the inter-molecular interactions, easing the task to scale to large systems. We verified that the two implementations result into the same trajectory up to numerical accuracy.

In both cases, the difference in free energy between the two states of the system, ΔF , is defined as:

	$\Delta F = F_{\lambda=1} - F_{\lambda=0} = \int_0^1 \langle \frac{\partial \Delta H(\lambda)}{\partial \lambda} \rangle_\lambda d\lambda$	(4.17)
--	--	--------

where $\langle \Delta H(\lambda) \rangle$ defining the ensemble average of the potential energy as a function of the perturbation parameter, λ . The derivation of Equation 4.17 is reported in Section 2.7.2.

The perturbation approach outlined in these equations, in general, is directly applicable. In the case of liquid samples, however, the molecular diffusion taking place at $\lambda = 1$ spoils the computation of $\langle \partial \Delta H(\lambda) / \partial \lambda \rangle$, since H_{harm} diverges quadratically with increasing distance of each molecule from its fixed minimum energy position.

Two strategies are used to deal with the problem of diffusion observed in liquid samples.

The first strategy consists in extending the integration up to $\bar{\lambda} < 1$ such that the integrand (and thus the integral) can be computed with a pre-set error bar. This of course represents an approximation but the results of the following sections show that λ is of the order of 0.995, thus accounting for most of the effects covered by thermodynamic integration. Moreover, we showed that the integration up to $\lambda = 0.995$ is able to remove the variations due to different local minima, validating the assumption that what remains above λ accounts for long-range diffusion, which is the same for systems made of the same solvent and at the same thermodynamic state (Sec. 4.4.1). These observations suggest that, despite the original aim of computing absolute free energies, limiting the integration up to $\lambda < 1$ is an approximation especially suitable to compute free energy differences, such as solvation energies.

The second strategy, in principle more rigorous and accurate, consists in modifying the purely harmonic approximation by making the harmonic Hamiltonian periodic, and thus bounded from above. Moreover, interchanges in the positions of local minimum between molecular pairs are introduced. As discussed in Section 4.4.2, these changes remove the singularity of the integrand at the cost of increased complexity

In both cases, starting from the minimized conformation, the harmonic reference was transformed into the actual system throughout an unphysical path discretized into ~ 16 individual windows. The spacing distance was optimized to increase the accuracy of the numerical integration of $\langle \partial \Delta H(\lambda) / \partial \lambda \rangle$ showing a singular behavior for λ approaching 1. Windows corresponding to $0.005 \leq \lambda \leq 0.95$ were simulated for a total of 1.6 ns. Due to the slow convergence of systems at λ equal to 0.99 and 0.995, robust statistics were collected for these points running production simulations of 3 and 5 ns, respectively. An additional point at λ equal to 0.9925 was added to improve the integral accuracy. The standard error was computed to assess the convergence of calculations.

In Table 4.4, the discretization of the unphysical path is summarized and the corresponding simulation time necessary to achieve convergence is also reported.

Table 4.4. Discretization of the unphysical path for thermodynamic integration.

λ	Spacing	Simulation time
0.005	0.005	1.6 ns
0.01	0.04	1.6 ns
0.05	0.05	1.6 ns
0.1 ... 0.9	0.1	1.6 ns
0.9	0.05	1.6 ns
0.95	0.04	1.6 ns
0.99	0.0025	3 ns
0.9925	0.0025	1.6 ns
0.995	-	5 ns

All the perturbation simulations were performed in parallel representing a modest computational effort for our small systems size (~ 80 h/sample on a single-core CPU).

4.3.7. Free energy decomposition

In order to calculate the free energy of a molecular system, we adopt the general strategy of computing the free energy difference between the fully interacting system and the corresponding reference state, for which the free energy is known. In our application, the so-called Debye model^{168, 146} was chosen as reference state. In Section 4.4.1, the validation of our reference is reported.

As already mentioned, the absolute free energy of a system is estimated including the quasi-harmonic, $F_{harm}(V)$, and the ideal, F_{id} , contribution arising from the integration over momenta, as well as the an-harmonic term, ΔF , computed by thermodynamic integration, which recovers the full an-harmonicity of the real system.

	$F = F_{harm}(V) + F_{id} + \Delta F$	(4.01)
--	---------------------------------------	--------

In addition to the absolute free energies, we compute hydration free energies (HFEs). HFE of a generic solute, A , in explicit solvent is defined as the difference between the free energy of the hydrated solute and the free energy referring to the bulk water and the pure solute, at the same thermodynamic conditions.

	$HFE_A = (F_{harm}(V) + F_{id} + \Delta F)_{A,wat} - [(F_{harm}(V) + F_{id} + \Delta F)_{bulk\ wat} + (F_{harm}(V) + F_{id} + \Delta F)_A]$	(4.18)
--	---	--------

In the following sections, the equations used to compute each contribution to the system free energy are reported and discussed.

4.3.7.1. Quasi-harmonic (QH) contribution

The quasi-harmonic (QH) term is obtained by first computing the harmonic free energy, $F_{harm}(V, T)$ on a mesh of equally spaced volumes at finite temperature, T . The computational data are interpolated by the Birch-Murnaghan equation of state (Eq. 4.12), whose minimum gives the equilibrium volume and free energy.

In QM, the expression of the volume-dependent harmonic free energy is:

	$F_{harm}(V, T) = -k_B T \ln Z = E_0 + \frac{\hbar\omega}{2} + k_B T \ln(1 - \exp[-\beta\hbar\omega])$	(4.19)
--	--	--------

where E_0 is the energy of the reference (minimum) configuration.

In the limit of $T \rightarrow 0$, $F_{harm}(V, T)$ equals the minimum of the potential energy plus the zero point energy. With increasing T , $F_{harm}(V, T)$ increases at first because of increasing potential energy, then it decreases because of the $-TS$ contribution.

In the classical limit, the volume-dependent harmonic free energy is defined as:

	$F_{\text{harm}}(V, T) = -k_B T \ln Z = E_0 - k_B T \sum_{i=1}^N \ln \left\{ \frac{k_B T}{\hbar \omega_i(V)} \right\} \quad (4.20)$	
--	---	--

In Figure 4.5, the behavior of the classical harmonic free energy as a function of the temperature is reported. It is similar to the quantum one, but shifted by $\sim E_{ZPE}$.

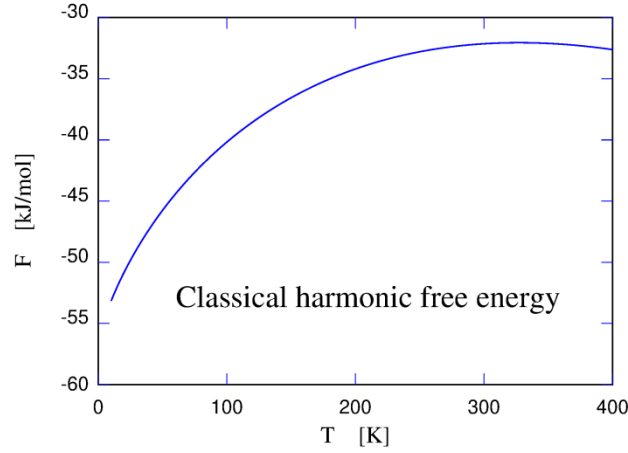


Figure 4.5. Classical harmonic free energy as a function of temperature. No liquid-vapor transition is observed in the harmonic approximation (HA).

In our computational scheme, we applied the classical statistical mechanics equation for the harmonic free energy as a function of the volume-dependent frequencies (Eq. 4.20).

4.3.7.2. Ideal contribution

The QH free energy estimate is complemented with the ideal contribution to the system free energy, accounting for the mass, volume, and temperature of the system.

In the classical limit, the ideal free energy for our fluid samples is computed as:

	$f_{CM} = k_B T \log \rho \Lambda_{CM}^3 \quad (4.21)$	
--	--	--

where $\rho = N/V$ and Λ_{CM} is the de Broglie wavelength of the system as a whole, whose mass M is concentrated at the center of mass, CM .

Λ_{CM} is defined as:

	$\Lambda_{CM} = \left(\frac{2\pi\beta\hbar^2}{M} \right)^{1/2} \quad (4.22)$	
--	---	--

For solid samples, the free energy of the center of mass is computed as:

	$f_{CM} = k_B T \log \frac{\Lambda_{CM}^3}{V}$	(4.23)
--	--	--------

The difference between Equation 4.21 and 4.23 arises from the fact that molecules are distinguishable in a solid sample.

Further details on the contribution of the ideal term to system free energy are reported in Section 2.1.3.

4.3.7.3. An-harmonic contribution from perturbation theory

The difference in free energy between the harmonic reference and the fully interacting system is computed by thermodynamic integration. The unphysical path transforming one state into another is discretized in a series of ~16 independent windows. For each window, the ensemble average of the partial derivative of the Hamiltonian as a function of the perturbation parameter λ is computed accordingly to the collected statistics as:

	$\Delta F = F_{\lambda=1} - F_{\lambda=0} = \int_0^1 \left\langle \frac{\partial \Delta H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda$	(4.24)
--	---	--------

where $\Delta H(\lambda)$ is the potential energy including the contributions of both real and reference systems.

Weights are assigned to each window based on its contribution to the trapezoidal integration. Finally, integrands are weighted and numerically integrated. Similarly to the requirements of umbrella sampling, windows need to be placed in such a way that probability distributions do overlap.

4.4. Results

As already mentioned, our computational scheme relies on the general strategy applied to compute the absolute free energy as the free energy difference between the fully interacting system and a representative reference state, whose free energy is known. We applied this strategy to condensed systems consisting of flexible water molecules and neutral organic solutes.

In our implementation, the so-called Debye model^{68, 146} was used as reference for both the fluid and the solid states. Through the Debye model, we defined an atomistic network model consisting of point particles (atoms) connected by harmonic springs determined by the Hessian matrix, which is computed at the atomic positions of local equilibrium.

In Section 4.4.1, the use of arbitrary local minima as source of reference coordinates for the Debye model is validated. Then, we discussed the difficulties related to the transformation of a quasi-harmonic system into a

fully interacting and diffusive liquid state, suggesting some possible solutions. Finally, the first application of the method is presented.

4.4.1. Validation of the reference system

Water has an incredibly complex phase diagram that at very low temperature is affected by QM effects. Nevertheless, we can assume that at ambient pressure the well-known Ih (hexagonal ice) structure is a deep minimum in the potential energy surface in the classical limit. This, however, is not a good reference system, since the crystal and the liquid are separated by a first-order phase transition. During a fast minimization procedure, disordered systems in explicit solvent are driven to one of a multitude of local minima.²⁰⁵ Our approach is based on the assumption that a disordered locally minimized configuration is a good source of reference coordinates giving a realistic description of the liquid system, since the liquid and the glass phases are connected by a glass transition that is continuous.

To validate our assumptions, we computed the main structural features of a system of 300 SPC/Fw water molecules from a relatively long trajectory in the NVE ensemble at 300 K.

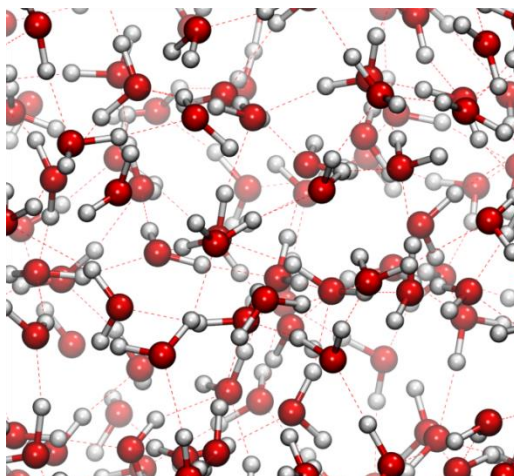


Figure 4.6. Snapshot of the water box. Through the Debye model, an atomistic network model consisting of point particles connected by harmonic springs determined by the Hessian matrix (covalent and hydrogen bonds) is defined.

Initially, the radial distribution functions corresponding to the pairs between H-H, O-H, and O-O (i.e. $g_{HH}(r)$, $g_{OH}(r)$, and $g_{OO}(r)$, respectively) were computed to evaluate the average distances among interacting atoms.

The radial distribution functions that we obtained are in good agreement with the experimental results (Fig. 4.7, Right).²⁰⁶⁻²⁰⁷ Starting from the top panel of Figure 4.7, Left, the radial distribution between hydrogen atoms, $g_{HH}(r)$, shows one peak at 1.7 Å corresponding to the intra-molecular H-H interaction. In the middle panel, the $g_{OH}(r)$ is reported showing the two peaks at 1.0 Å and 1.7 Å representing the intra-molecular and the inter-molecular O-H interactions, respectively. In the bottom panel, the radial distribution function between oxygen atoms, $g_{OO}(r)$, shows one single sharp peak at 2.8 Å corresponding to the distance between oxygen atoms of interacting water molecules.

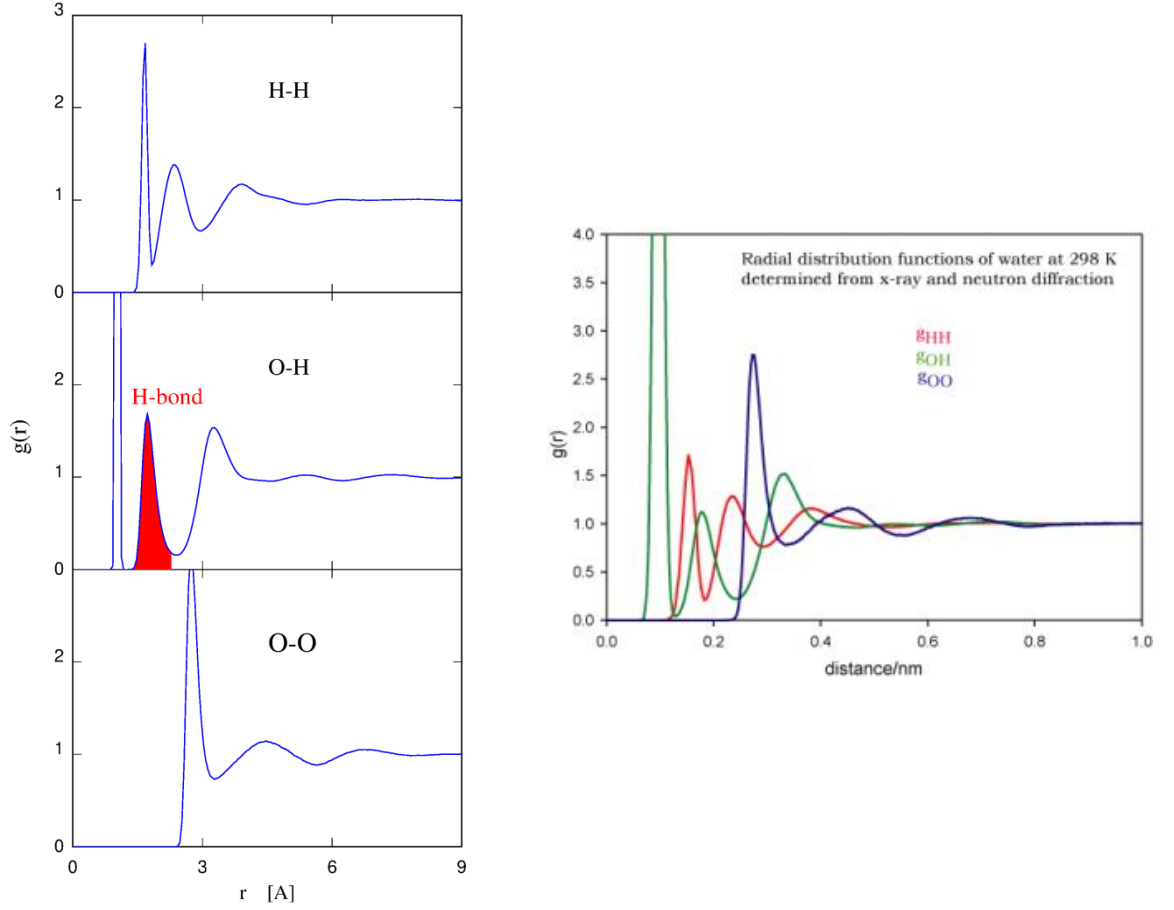


Figure 4.7. (Left) Radial distribution functions obtained for the system of 300 water molecules. From top to bottom panels, the radial distribution functions are reported corresponding to the interactions between H-H, O-H, and O-O, respectively. (Right) Experimental radial distribution functions of liquid water.²⁰⁷

The dynamics of 300 SPC/Fw¹⁹⁷ water molecules at 300 K was validated computing the diffusion coefficient from the time dependence of the average square distance over all oxygen atoms, returning the mean square displacement, $MSD(t)$, defined as:

	$MSD(t) = \langle \Delta \mathbf{r}(t)^2 \rangle = \langle (\mathbf{r}(t + t_0) - \mathbf{r}(t_0))^2 \rangle_{t_0}$	(4.25)
--	---	--------

where $\langle \dots \rangle_{t_0}$ indicates the average over the initial time (and configuration) along the trajectory. In our computations, this is obtained by a running average over from 100 initial configurations. The result is displayed in Figure 4.8 over 200 ps out of a 400 ps trajectory.

When $MSD(t)$ reaches the linear regime, the slope of $MSD(t)$ relates to the self-diffusion constant, D , accordingly to Einstein's relationship:

	$\lim_{t \rightarrow \infty} \frac{MSD(t)}{t} = 2dD$	(4.26)
--	--	--------

where d is the system dimensionality.

Thus, the self-diffusion coefficient of the SPC/Fw water model results equal to $2.250\text{E-}05\text{ cm}^2/\text{s}$, which is in excellent agreement with the experimental value of $2.3\text{E-}05\text{ cm}^2/\text{s}$ reported by Andanson et al.²⁰⁸

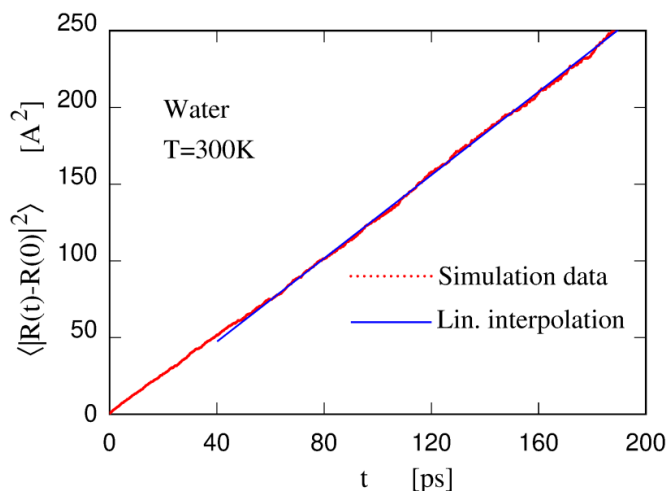


Figure 4.8. Mean square displacement computed on a system of 300 SPC/Fw¹⁹⁷ water molecules at 300 K. The simulation data and the linear interpolation are reported in red and blue, respectively.

The H-bonding network characterizing the liquid phase of water was evaluated computing the number of hydrogen bonds established among 300 interacting water molecules. To assess the presence of one H-bond, the distance between the oxygen atoms of interacting molecules was considered, as well as the angle centered on the hydrogen atom involved in both covalent and H-bond interactions.

In the system of 300 water molecules, taking $d(O - O) < 3.2\text{ Å}$ and $\theta(OH - -O) > 140^\circ$, ~520 H-bonds were identified out of a total of 600 possible nearest-neighbor interactions, suggesting the presence of a non-negligible number of vacant H-bonds (Fig. 4.9). The computational result is in fair agreement with experimental estimates.²⁰⁹

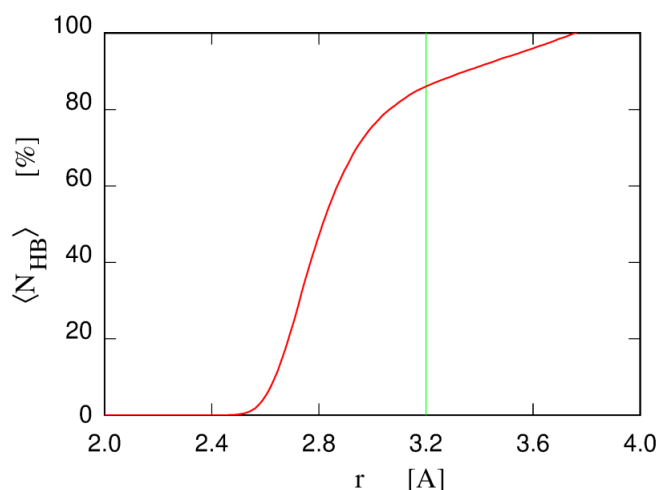


Figure 4.9. Probability distribution of distance between H-bonded oxygen atom pairs. 100% corresponds to two hydrogen bonds per water molecule. The hydrogen bonds is defined purely in terms of geometry, as a pair of oxygen atoms with an hydrogen in between, forming an angle $OH \cdots O > 140^\circ$. Results averaged over 100 ps.

The intrinsic dynamics of bulk water was evaluated by computing the vibrational frequencies from the Hessian matrix and reconstructing its vibrational density of state, vDOS (Fig. 4.10, Left). In the presence of a solute, the projected density of states (pDOS) was considered (Fig. 4.10, Right). We obtained a good reproduction of the experimental vibrations of liquid water in both vDOS and pDOS.²¹⁰ In particular, high frequency O-H stretching and bending are identified at $\sim 3500\text{ cm}^{-1}$ and $\sim 1600\text{ cm}^{-1}$, respectively. The resulting DOS at low frequencies (i.e. $\omega < 1000\text{ cm}^{-1}$) due to inter-molecular interactions is affected by the disorder of the system giving origin to an excess of localized states. Because of their low frequencies, inter-molecular modes are those that mostly contribute to the total entropy and free energy.

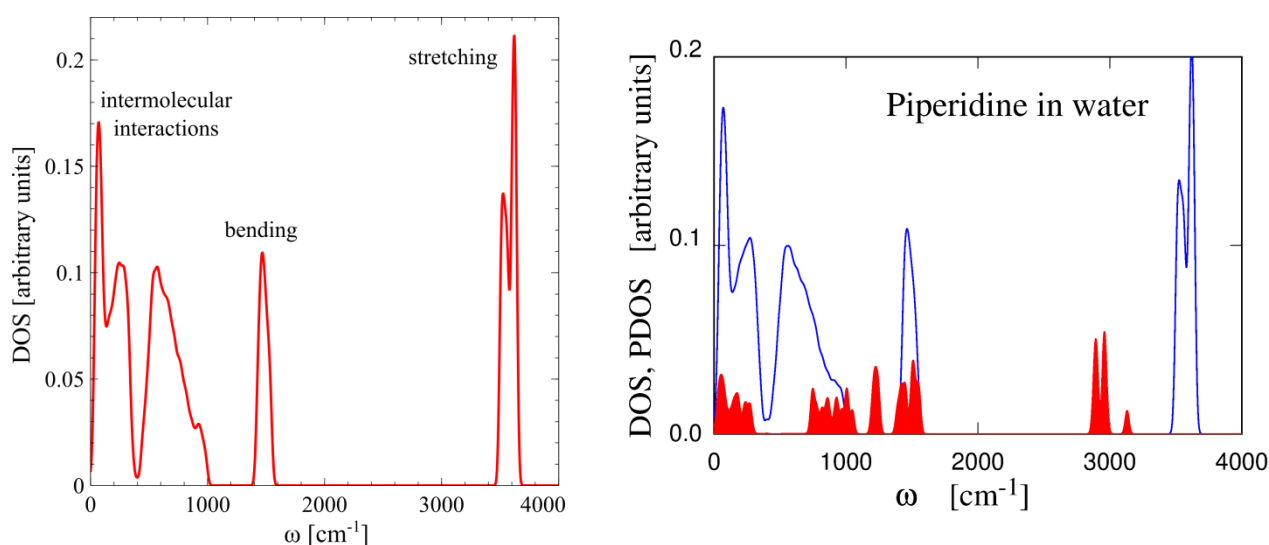


Figure 4.10. (Left) Vibrational density of states (vDOS) computed on a system of 300 SPC/Fw water molecules. (Right) Projected on piperidine density of state (pDOS) computed for the system including one molecule of piperidine in the neutral state solvated in 300 water molecules. The DOS of the solute is red colored and intensified to be visible. Some smoothing applied.

The participation ratio was applied to localize the low frequency modes.⁶⁹ The excess of localized modes is identified by comparing the vDOS of liquid water and the vDOS of hexagonal ice Ih, whose modes are delocalized. In ordinary ice Ih, oxygen atoms sit on a lattice of hexagonal symmetry whose space group is $P6_3/mmc$. Each hydrogen atom lies on the axis joining two oxygen atoms of neighboring molecules, with four other water molecules to form a tetrahedron. In ice Ih, protons can adopt many different arrangements between oxygen atoms. Proton-disordered configurations in ice Ih originate the zero point entropy, which is related to the fact that at 0 K, when no thermal agitation exists any longer, there are still a number of possible configurations for a given ice crystal. Linus Pauling estimated the number of possible hydrogen-bonded configurations for an ice crystal at 0 K²¹¹ under some assumptions regarding the ice structure that are also known as the Bernal-Fowler (BF) ice rules.²¹²

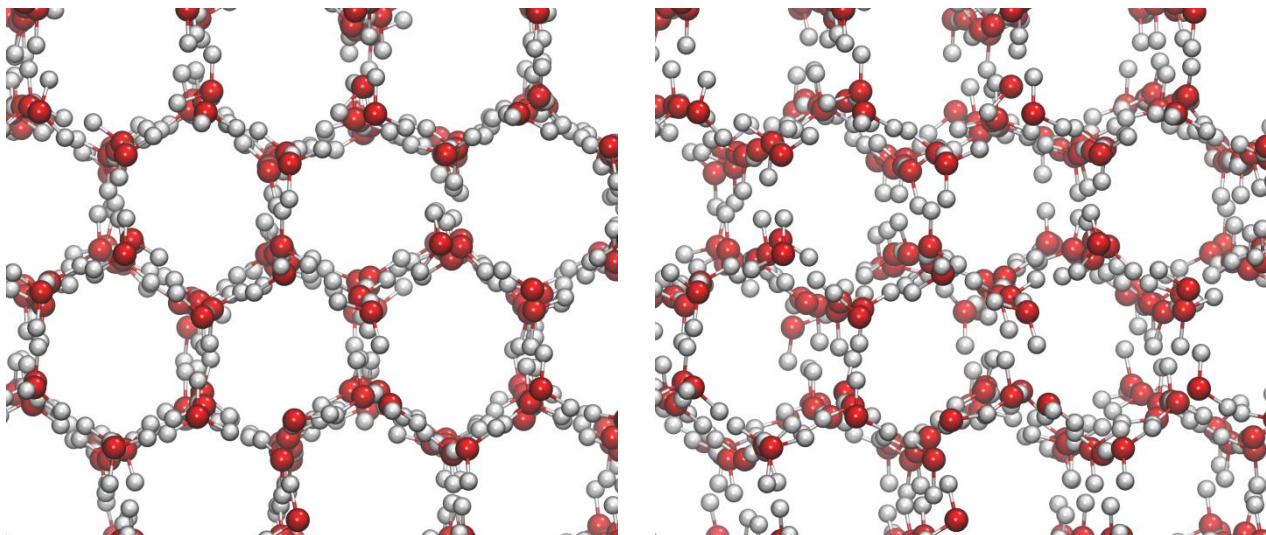


Figure 4.11. Snapshots representing one proton-disordered configuration of ice Ih.

In our analysis, a model consisting of 1,024 molecules of ice Ih (i.e. 3,072 particles) in an orthorhombic box was built. Proton-disordered configurations were produced by running 1 ns MD simulation in the NVE ensemble, keeping the temperature under the melting point characterizing the water model, which was determined to be around 200 K for the rigid SPC.²¹³ In proton-disordered configurations, hydrogen atoms were allowed to modify the overall hydrogen bonding network without distorting the hexagonal geometry of the system (Fig. 4.11). The disordered ice Ih configuration was minimized by quenched MD and the corresponding harmonic Hamiltonian was computed and diagonalized. In Figure 4.12, the resulting vDOS of liquid water and ice Ih are compared.

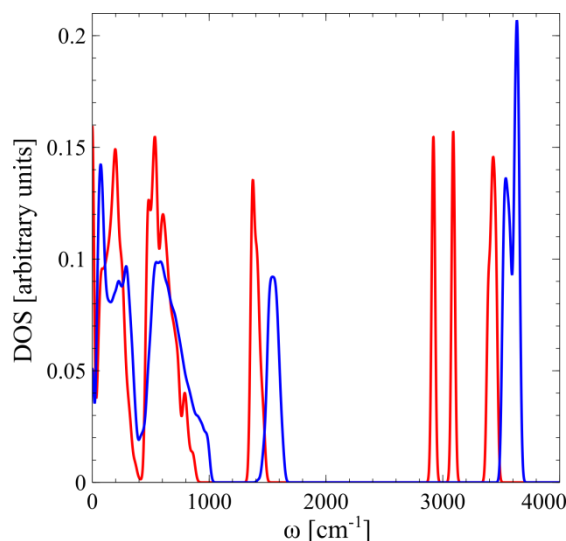


Figure 4.12. Comparison of the vDOS computed for liquid water (blue) and ice Ih (red). The two high-frequency bending peaks are related to the symmetric and asymmetric stretching of water molecules. Some smoothing applied.

Finally, monitoring the potential energy of the system while continuously lowering the temperature, we verified that the potential energy changes continuously across the range of explored temperatures, suggesting

the presence of a continuous glass phase transition where the arbitrary local minima represent the disordered system in glass phase (Fig. 4.13).

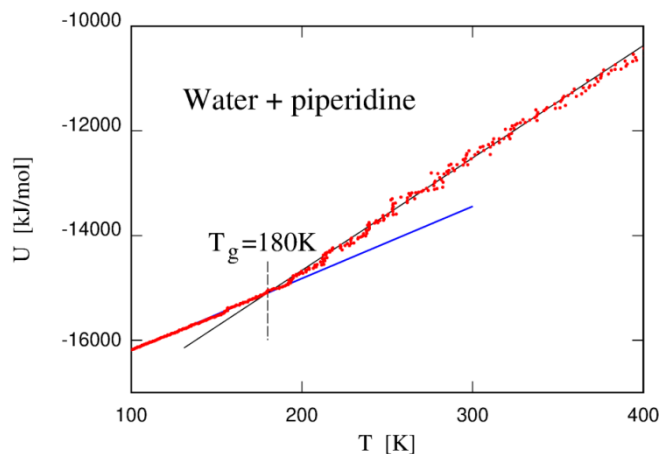


Figure 4.13. Glass phase transition computed on the system of piperidine solvated in 300 SPC/Fw water molecules. The glass transition temperature, T_g , is identified by the intersection of the two linear curves fitting the potential energy as a function of the temperature related to the liquid and glass phases (higher and lower temperatures, respectively).

We verified that the probability distribution for the potential energies and free energies associated to various local minima explored during the equilibration dynamics can be quite broad (Fig. 4.14). As a consequence, arbitrary minima of the same system can have significantly different potential energy.

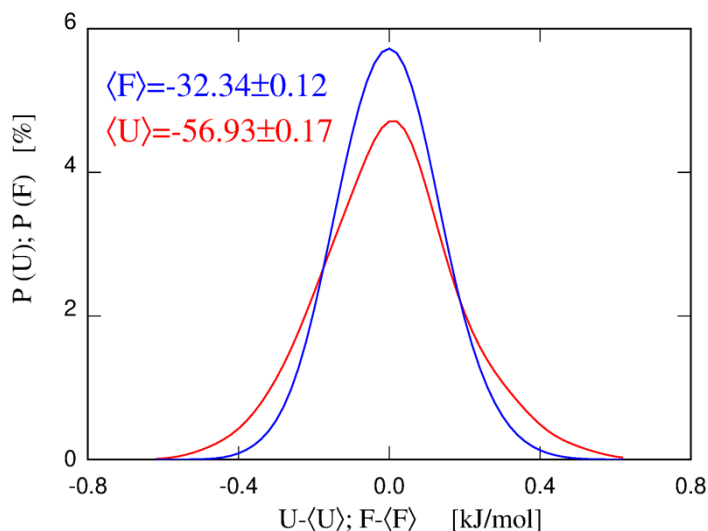


Figure 4.14. Probability distribution for the free energy (blue) and potential energy (red) per molecule computed on energy minimized configurations. The curve has been obtained from 400 distinct minima determined by minimizing the energy of configurations selected every 10,000 MD steps at 300 K.

We verified that the potential energy difference among arbitrary local minima is compensated by perturbation dynamics. To this aim, we estimated the free energy of 300 SPC/Fw water molecules starting from two configurations of local minima characterized by significantly different potential energies ($\Delta U = 185$ kJ/mol). Upon the estimate of the QH contributions to the free energy, the free energies of the two systems differed by 156 kJ/mol. Adding the TI term up to $\lambda = 0.99$, their free energy difference went down to 3 kJ/mol (or 0.01 kJ/mol for the single water molecule). Figure 4.15 shows how the TI compensates the

free energy difference between the two local minima. In particular, it suggests that the compensation occurs at $\lambda < 0.3$ and $0.8 \leq \lambda \leq 0.99$, whereas for $\lambda > 0.99$, the statistical error increases rapidly and unpredictably. The intermediate range corresponding to $0.3 \leq \lambda < 0.8$ seems not to significantly contribute to the compensation of the free energy difference between the two arbitrary local minima.

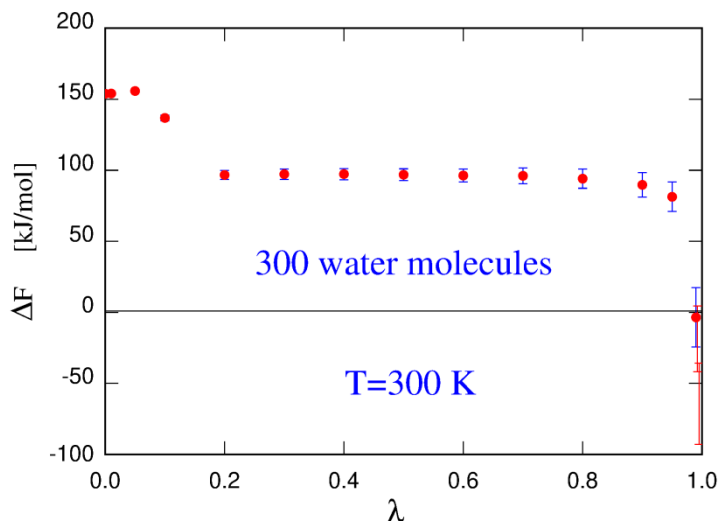


Figure 4.15. Effect of the thermodynamic integration (TI) contribution on the free energy difference between two independent local minima.

In conclusion, we verified that using the flexible SPC/Fw¹⁹⁷ water model, the main structural and dynamical features of liquid water can be reproduced in fair agreement with the corresponding experimental results. We verified that an arbitrary local minimum is a good source of reference coordinates for the liquid system being connected by a continuous (glass) phase transition to the liquid state. Moreover, we showed that by adding the TI contribution up to $\lambda = 0.99$, one can compensate the free energy differences shown by arbitrary local minima. Consequently, limiting the integration up to $\lambda < 1.0$ is an approximated approach suitable to the estimate of free energy differences, such as hydration free energies.

4.4.2. Recovering the full an-harmonicity

As previously mentioned, the potential energy surface of water system is rugged because of the multitude of nearly equivalent hydrogen bonding conformations. Upon quenching, a system can be driven to different arbitrary minima showing non-negligible energy differences that have to be compensated by perturbation theory. This point is particularly relevant in our approach, where the systems including the hydrated solute, the bulk water, and the pure solute at normal conditions, are treated independently. Thus, energetically unrelated arbitrary minima are used as sources of reference coordinates to compute and diagonalize the Hessian matrix for these three independent samples.

The first partial compensation of these differences comes at the QH level. Disordered configurations of higher energy tend to have lower harmonic frequencies, reflecting their lower stability. Lower frequencies, in turn, correspond to higher entropy and to a faster decrease of free energy within increasing temperature, thus

contributing to the equalization of the free energy estimated from different local minima. Most of the compensation, however, is achieved by the TI step, which, moreover, recovers the full an-harmonicity of the fully-interacting system described by the force field, $[U_{ff}(\mathbf{r}^N)]$.

To this aim, the potential energy as a function of the perturbation parameter, $U(\lambda)$, was defined as linear interpolation between the two states, setting $\lambda = 1$ for the description of the real system and $\lambda = 0$ for the harmonic one.

	$\Delta H(\lambda) = U(\lambda) = \lambda U_{ff} + (1 - \lambda) U_{harm}$	(4.27)
--	--	--------

Defining $\Delta H(\lambda) = H(\lambda) - H_{harm}(0)$, the free energy difference is then computed as:

	$\Delta F = F_{\lambda=1} - F_{\lambda=0} = \int_0^1 \left\langle \frac{\partial \Delta H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda = \int_0^1 \langle U(\lambda) - U_{harm} \rangle_{\lambda} d\lambda$	(4.28)
--	---	--------

where $\langle \dots \rangle_{\lambda}$ means that the average is estimated on the trajectory computed with $U(\lambda)$. The definition of $\Delta H(\lambda)$ can be more general than Equation 4.27, including a non-linear λ -dependence in $U(\lambda)$. This point will be discussed in Section 4.4.2.1. It is important to remark that this equation is exact, and $F_{\lambda=1} = F_{\lambda=0} + \Delta F$ is the free energy of the fully-interacting system described by the force field.

4.4.2.1. Addressing the molecular diffusion of fluid samples

As already stated in Section 2.9 and discussed in several papers, the major difficulty in the thermodynamic integration step is the onset of molecular diffusion in fluid samples. With increasing λ , in particular, atoms move more and more away from their minimum energy positions, since the restraining forces decrease and tend to zero in the $\lambda \rightarrow 1$ limit. The growth of the mean square displacement $\langle R^2 \rangle$ with increasing λ is shown in Figure 4.16, Left, documenting the increasing exploration of the phase space by molecules while decreasing the harmonic restraints. The leading term in the divergence of $\langle R^2 \rangle$ for $\lambda \rightarrow 1$ could in fact be predicted fairly easily. The unbound growth of $\langle R^2 \rangle$ implies the divergence of H_{harm} to $+\infty$, and thus the divergence of the TI integrand $\langle \partial \Delta H(\lambda) / \partial \lambda \rangle_{\lambda}$ towards $-\infty$, as shown in Figure 4.16, Right. Notice that the TI integral remains finite, since it represents the difference of two free energies ($F_{\lambda=1}$ and $F_{\lambda=0}$), which are both finite.

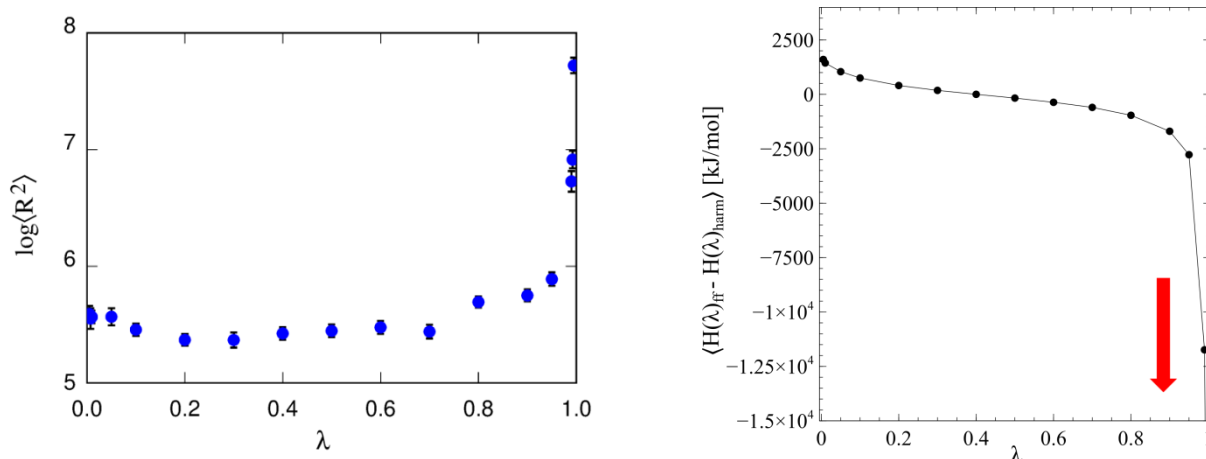


Figure 4.16. (Left) The square of the radius explored by the system at different values of the perturbation parameter, λ . The divergence at $\lambda=1$ suggests that particles diffuse for $\lambda \rightarrow 1$. Semi-logarithmic scale. (Right) Linear interpolation between reference and fully interacting states. The integrand values are reported on the y-axis. The integrand diverges to $-\infty$ for $\lambda \rightarrow 1$.

To overcome this problem, it has been proposed to limit the harmonic Hamiltonian to a finite value, since there is still much freedom left in the choice of the reference model. To this aim, we first make the harmonic restraint periodic in space, with the periodicity L of the simulation cell. More in detail, let us consider the case of a pure water sample and water molecules that perform oscillations around their minimum energy position. With increasing λ , every water molecule could move closer to a periodic replica of the minimum energy position than to the original one. In such a case, the new implementation adopts the replica as the origin of its restraint. Notice that this modification will not affect the simulation at low λ (i.e. $\lambda < 0.9$) since in that regime no water molecule will move farther than a nearest neighbor distance, and certainly much less than half the box size. Therefore, this change will not spoil the harmonic reference state at $\lambda = 0$.

This is not as simple as making periodic the force field interaction among particles, which, in any case, is supposed to be negligible at distances comparable to $L/2$, where L is the box size. The harmonic potential and its derivative grow monotonically in every direction, and at $L/2$ they are usually large (Fig. 4.17, red line). More importantly, the force of the restraint suddenly changes sign at $L/2$, making the simulation unstable.

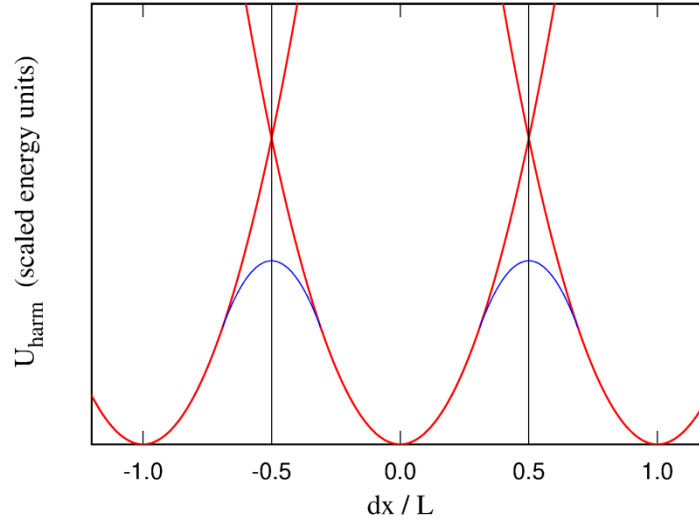


Figure 4.17. Behavior of the energy profile of the restraint across the simulation cell. The harmonic potential as a function of the linear separation dx is red colored. The energy of the harmonic restraint across the simulation cell after the coordinate transformation reported in Equation 4.29 is blue colored.

To achieve periodicity without introducing discontinuities in the forces, we replace the linear separation dx of molecules and restraining centers, with a smoothed version dx' that bends at the boundary of the simulation cell:

	$\frac{dx'}{L} = \begin{cases} dx/L & \text{if } dx/L \leq 0.5 \\ 0.475 - 10(dx/L - 0.5)^2 & \text{if } dx/L \geq 0.45 \\ -0.475 + 10(dx/L + 0.5)^2 & \text{if } dx/L \leq -0.45 \end{cases}$	(4.29)
--	---	--------

where L is the box side. This transformation is illustrated in the top panel of Figure 4.18. With this coordinate transformation, the energy profile of the restraint across the simulation cell changes into the blue line reported in Figure 4.17. Strictly speaking, this modification is already enough to remove the divergence of the TI integrand, although this step is not sufficient to make the computation fully manageable, since forces still changes rapidly across the cell boundary. This is easily verified considering that forces from the quadratic restraint are proportional to $d dx'/d dx$, as shown in the lower panel of Figure 4.18.

The derivative of this effective separation dx' with respect to the true separation dx turns out to be:

	$\frac{d dx'}{d dx} = \begin{cases} dx'/dx & \text{if } dx/L \leq 0.5 \\ -10(dx/L - 0.5) & \text{if } dx/L \geq 0.45 \\ 10(dx/L + 0.5) & \text{if } dx/L \leq -0.45 \end{cases}$	(4.30)
--	--	--------

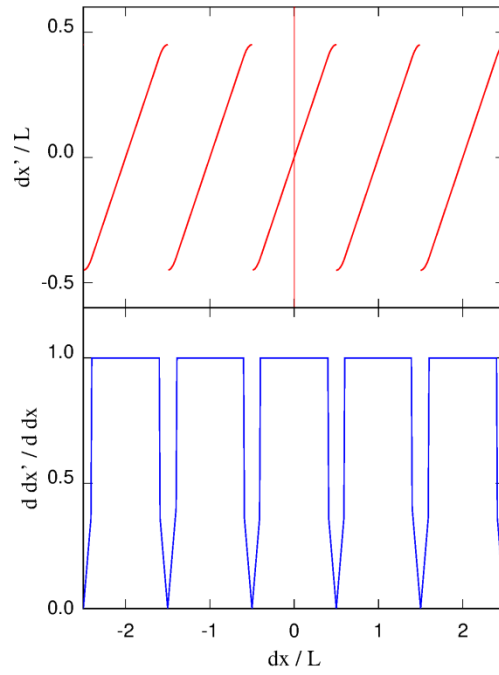


Figure 4.18. (Top panel) Illustration of the transformation from dx to dx' . (Lower panel) Behavior of $d(dx')/d(dx)$ as a function of dx/L . See Equations 4.29 and 4.30.

To ease this problem, we change the sign of the displacement dx whenever dx is in an odd-numbered box, counting boxes from box 0, which corresponds to the original one. The correspondence of true distance dx , fictitious distance dx' , and derivative ($d dx'/d dx$) is shown in Figure 4.19.

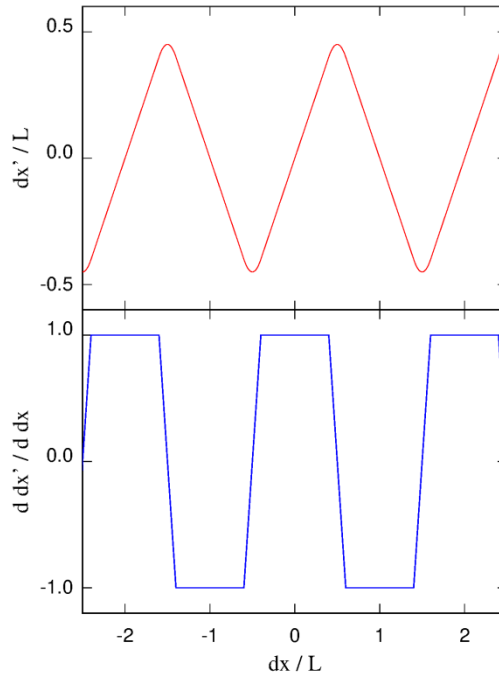


Figure 4.19. Behavior of dx'/L and $d(dx')/d(dx)$ as a function of dx/L after changing the sign of the displacement dx .

Once these changes have been made, the divergence of the integrand is no longer a strict concern. However, the integrand still has a sharp rise for $\lambda \sim 1$, and we explored a way to overcome also this problem.

A promising strategy has been proposed by Tyka and coworkers,¹³⁶ suggesting that molecules could be interchanged to minimize their harmonic energy. This would greatly limit the highest value of the harmonic energy and thus of the integrand. This advantage, however, comes at the cost of searching the optimal mapping of molecules and reference minima for the computation of the harmonic Hamiltonian. The number of ways each molecule can be assigned to a minimum grows with the system size as the factorial of the number of molecules. Fortunately, smarter algorithms have been proposed to solve this so-called linear assignment problem,²¹⁴ reducing the scaling to a polynomial (fifth degree) power of the number of molecules. However, even a fifth degree algorithm looks too expensive for any system exceeding a few tens of molecules.

Hence, we resorted to a Monte Carlo type of approach, as already done, for instance, by Berryman and Schilling.²¹⁵ At regular intervals during the simulation, we attempt the exchange of a pair of molecules. With the same frequency, we also attempted the flip of the hydrogens on the same water molecule, H1 and H2. The two processes are both needed to achieve the optimal decrease of the integrand. Such an exchange will not cause a difference in the energy and forces due to the original force field, but it implies a change ΔE of the harmonic energy, whose estimate is inexpensive. Since the energy variations usually are much higher than $k_B T$, we use in fact a zero temperature Monte Carlo, accepting the swap whenever $\Delta E < 0$.

In practice, with the simple λ interpolation used at this stage, exchanges occur only at the highest λ values, such as $\lambda > 0.95$ and do not need to be sampled at lower λ , as could already be inferred from the plot of the average displacement in Figure 4.16, Left. To account for the huge number of possible associations of molecules and minima, 100 interchanges are attempted at each MD step at $\lambda > 0.95$, with an acceptance probability of the order 10^{-4} per attempt for $\lambda = 0.995$. We observe that exploiting the system periodicity and introducing pair swaps is done more easily and naturally using the implementation in the original Cartesian coordinates instead of the displacement along eigenvectors (Sec. 4.3.6). Also this modification of the original algorithm preserves the harmonic Hamiltonian and free energy for $\lambda \rightarrow 0$.

This pragmatic swap algorithm has the effect of reducing the integrand in the TI step at an acceptable computational cost. For instance in the case of 300 water molecules at 300 K, the integrand at $\lambda = 0.995$ decreases from $\langle \partial H / \partial \lambda \rangle = -56,426$ kJ/mol to $\langle \partial H / \partial \lambda \rangle = -24,718$ kJ/mol (with H1 and H2 flips). However, each swap discontinuously decreases the harmonic energy by an amount ΔE that often is not negligible. To overcome this issue, the amount of energy ΔE is transformed into kinetic energy, rescaling the velocity of all particles. In this way and considering also the effect of the Langevin thermostat, the simulation runs smoothly for millions of steps, conserving, on average, energy and temperature. The rigorous identification of the ensemble in which the simulation is performed, however, is somewhat blurred, and for this reason we quote as final results those obtained with the simpler method without exchanges.

We point out that this difficulty is strictly the consequence of the statistical swap approach, finding water pairs to be exchanged far after their swap became favorable.

In principle, the potential energy surface would be continuous provided one could identify at every step the permutation of water molecules to restraints assignment that minimizes the harmonic energy term. The task of identifying the optimal solution of this linear assignment problem can be carried out following the exact algorithm,²¹⁶ whose implementation, however, requires a number of operations scaling unfavorably with the system size (N^5). Since we aim at large systems, this systematic approach is of limited usage. In a pragmatic way, to approach the requested *adiabatic* surface condition, we first increased the frequency of interchanges at each MD step, while providing the systematic search of the optimal assignment in terms of minimization of the harmonic potential energy. This approach was motivated by the following intuitive considerations that, however, turn out to be faulty: for the configuration of local energy minimum, the optimal assignment is the identity permutation. Then, by a sort of continuity in the space of operators, we expected that the first favorable permutations able to decrease the harmonic energy would be the pair exchanges. Hence, performing them exhaustively ($[N \times (N - 1)/2]$ distinct exchanges for N water molecules) at each step would solve the problem. This assumption, however, is not correct, since the first permutation decreasing the harmonic energy could be more complex than a pair exchange. A pictorial scheme is reported in Figure 4.20. Therefore, we concluded that only a more comprehensive search for the optimal permutation would achieve the goal to a continuous potential energy surface for the mixed Hamiltonian.

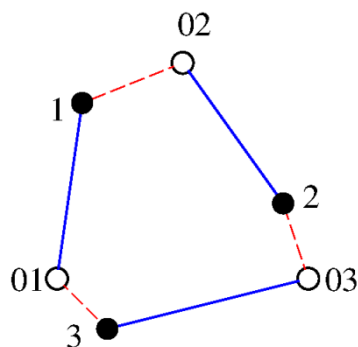


Figure 4.20. The scheme represents three particles (1, 2, 3) harmonically restrained to their equilibrium positions (01, 02, and 03, respectively). The permutation $123 \rightarrow 321$ decreases the sum of the squares of the restraint distances. Every binary interchange, instead, increases the corresponding sum.

Attempting the permutations of n molecules at a time is not much more expensive than swapping pairs, but the number of these different assignments equal to $N!/(N - n)!$, is staggering as soon as n is beyond the few molecules. We endeavored to identify a restricted set of irreducible permutations, exploiting known properties of representations of the permutation group. In particular, for each n we need the subset of permutations whose representation has character 0. The aim is to show that the number of these irreducible permutations is a portion of the total number decreasing with n , and testing their total number for all n requires less effort than the accepted systematic solution. The proof of these statements is not trivial, and we

are still working of this development. Without a convincing solution to the problem of identifying the optimal assignment, we think that the loss of a clear identification of the statistical mechanics ensemble caused by the MC swap of molecules is a very severe problem, limiting the practical value of the method especially for large systems.

To overcome this problem we explored an alternative approach, close in spirit and in practice to the one we already used.

In this approach the role of water diffusion for $\lambda \rightarrow 1$ is limited by performing the perturbation dynamics from the harmonic state to the fully interacting system at low temperature (100 K in our application), located below the glass temperature of water that in our model takes place at $T \sim 180$ K (Fig. 4.13). Then, we add this contribution to the QH term to obtain the absolute free energy of the system at this low energy state. Finally, we quantify the free energy that the system gains increasing the temperature up to 300 K, by simulating the fully interacting system modeled by the force field at intermediate temperatures. Since no discontinuous melting transition separates the two states, the computation of the free energy variation along the path from $T_1 = 100$ K and $T_2 = 300$ K is conceptually straightforward, although, perhaps, computationally demanding.

In the following section, we report the derivation of Equation 4.37, which was applied in this approach to compute the system free energy at 300 K.

We started from the definition of the excess free energy as a function of the configurational partition function:

	$e^{-\beta F_{ex}(\beta)} = \int d\mathbf{r}^N e^{-\beta U(\mathbf{r})}$	(4.31)
--	--	--------

where $\beta = 1/k_B T$ and $F_{ex}(\beta)$ is the excess free energy.

Taking the derivative of Equation 4.31 with respect to β , we obtained:

	$-\left[\beta \frac{\partial F_{ex}}{\partial \beta} + F_{ex}\right] e^{-\beta F_{ex}(\beta)} = - \int d\mathbf{r}^N U e^{-\beta U(\mathbf{r})}$	(4.32)
--	--	--------

With minor transformations, we derived the expression defining the ensemble average of the potential energy as a function of β :

	$\beta \frac{\partial F_{ex}}{\partial \beta} + F_{ex} = \frac{\int d\mathbf{r}^N U e^{-\beta U(\mathbf{r})}}{\int e^{-\beta U(\mathbf{r})} d\mathbf{r}^N} = \langle U(\beta) \rangle$	(4.33)
--	--	--------

Hence,

	$\frac{\partial}{\partial \beta} (\beta F_{ex}) = \langle U(\beta) \rangle$	(4.34)
--	---	--------

By integrating both terms, we derived Equation 4.35:

	$\beta_2 F_{ex}(\beta_2) - \beta_1 F_{ex}(\beta_1) = \int_{\beta_1}^{\beta_2} \langle U(\beta) \rangle d\beta$	(4.35)
--	--	--------

Equation 4.35 was expressed as a function of the system temperature, T , with $\beta = 1/k_B T$, $d\beta = -k_B dT / (k_B T)^2$, and T_1, T_2 the initial and target temperatures of the system:

	$\beta_2 F_{ex}(\beta_2) = \beta_1 F_{ex}(\beta_1) - k_B \int_{T_1}^{T_2} \frac{\langle U(T) \rangle dT}{(k_B T)^2}$	(4.36)
--	--	--------

and

	$F_{ex}(T_2) = \frac{T_2}{T_1} F_{ex}(T_1) - k_B T_2 \int_{T_1}^{T_2} \frac{\langle U(T) \rangle dT}{(k_B T)^2}$	4.37
--	--	------

Equation 4.37 was then used to compute the system free energy at the final target temperature, T_2 .

By applying this alternative approach to compute free energies of systems in explicit solvent, we expect to limit the contribution of the integrand for $\lambda \rightarrow 1$ by performing the thermodynamic integration at low temperature (T_1) and to recover the free energy of the system at the target temperature, T_2 , by perturbing the fully interacting system between T_1 and T_2 . We emphasize that T_1 has to be below the glass temperature of water, before the onset of fast diffusion that could prevent the accurate computation of the $F_{ex}(T_1)$ term. We tested this option to one representative compound of our series (i.e. propionic acid) obtaining very promising results although we need to improve statistics, suggesting that the integrand for $\lambda \rightarrow 1$ can be computed fairly accurately despite a very weak divergence of the integrand, due to the low residual diffusion in the glass state. As already discussed this divergence is integrable, and in this case its contribution is small.

A second remark is in order. The sharp raise of the integrand at $\lambda \sim 1$ apparently reflects the discontinuous onset of diffusion at $\lambda = 1$. Making the change more continuous, enhancing the mean square displacement $\langle R^2 \rangle$ already at lower λ , would ease this problem making the TI estimate more accurate and reliable. The fast change due to the linear interpolation between reference and actual states can be limited by modifying the interpolation itself.²¹⁵ For example, the elongation, dx , of each particle can be reduced by a linear term depending on the perturbation parameter, λ . Thus, the parabolic behavior for small elongation happening at

lower λ values would be unaffected, whereas the diffusing particles would be limited by the linear trend of the potential energy walls.

An example of alternative interpolation of the end states is reported in Equation 4.38 and represented in Figure 4.21. Also in this case, the harmonic case is recovered in the $\lambda \rightarrow 0$ limit.

	$\tilde{U}_{harm}(\lambda) = \frac{1}{2} \omega^2 \left(\frac{dx^2}{1 + k\lambda dx } \right)$	(4.38)
--	--	--------

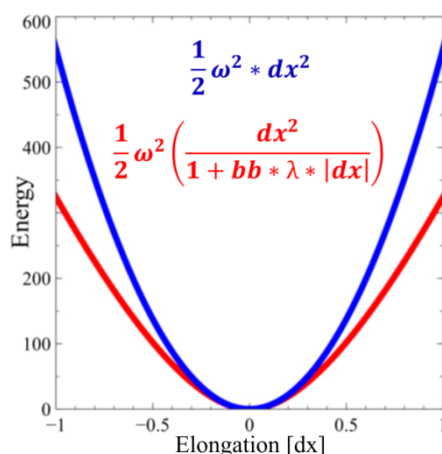


Figure 4.21. Beyond the Debye model. Energy behavior with respect to the particle elongation, dx . The behavior resulting from the linear and alternative interpolations between reference and actual states are blue and red colored, respectively.

These modifications are more easily introduced in the implementation using the displacement along eigenvectors, and, in fact, this has provided the motivation to develop that implementation.

Unfortunately, the improvement provided by this recipe is modest. Moreover, a sizeable effort is required to maximize the effect while preserving the harmonic character of the Hamiltonian at low λ and thus retaining the analytical reference free energy.

4.4.3. First test application: Hydration free energy (HFE) estimates

As first test application of the method, hydration free energy (HFE) estimates were computed for the selection of globally neutral small molecules reported in Table 4.2. Solvation and desolvation drive many important biological and chemical processes, including binding, adsorption, protein-ligand and protein-protein interactions, membrane formation, and folding. Hence, modeling the solvation component is important in computational biology and chemistry.

Thus, for a generic solute, A, the HFE is computed as:

	$HFE_A = (F_H(V) + F_{id} + \Delta F_{TI})_{A,wat} - [(F_H(V) + F_{id} + \Delta F_{TI})_{wat} + (F_H(V) + F_{id} + \Delta F_{TI})_A]$	(4.39)
--	---	--------

where $F_H(V)$ is the quasi-harmonic (QH) free energy, F_{id} is the ideal term, and ΔF_{TI} is the perturbative free energy computed by thermodynamic integration converting the harmonic Debye model into the fully interacting system described by the force field.

By computing HFE, which is in fact a free energy difference, the critical issues presented in the previous sections have been overcome. In the difference, the effect of diffusion of water molecules was supposed to cancel out subtracting the bulk water contribution to the term related to the solvated solute. This is qualitatively confirmed by Figure 4.21 showing the behavior of the difference reported in Equation 4.40:

	$\Delta I(\lambda) = \int_0^\lambda \frac{dF^{k+w}(\lambda')}{d\lambda'} d\lambda' - \int_0^\lambda \frac{dF^w(\lambda')}{d\lambda'} d\lambda'$	(4.40)
--	---	--------

where the first term refers to the system of water-solute (in this case, ketoprofen, but it is extendable to every solute molecule) and the second one to the system of bulk water.

Notice that both integrals in Figure 4.22 diverge at $\lambda = 1$ whereas the difference tends to converge.

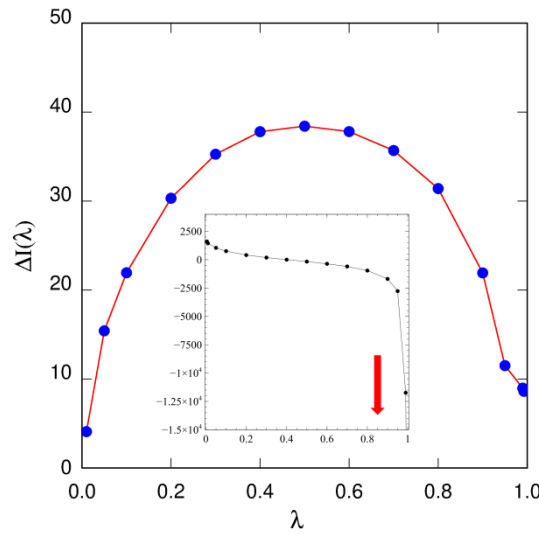


Figure 4.22. Difference in the thermodynamic integration contributions of (ketoprofen in water) and (pure water). See Equation 4.40 for the definition of $\Delta I(\lambda)$. The inset shows the behavior of the TI integrand as a function of λ for the pure water system.

Moreover, the integration was limited at λ values allowing particles exploring the conformational space around the local minimum without diffusing. More precisely, we considered the range $0.005 \leq \lambda \leq 0.995$.

First, we computed the free energy contribution of one solute molecule hydrated in a box of 300 SPC/Fw¹⁹⁷ water molecules by summing the absolute ($F_H(V)$ and F_{id}) and perturbative (ΔF) terms determined following the computational schemes presented in the previous sections.

Similarly, the free energy contribution due to bulk water was quantified. The calculation was performed on the system consisting of 300 SPC/Fw¹⁹⁷ water molecules obtaining an estimate in excellent agreement with

our theoretical benchmark computed by Habershon and Manolopoulos¹⁴⁶ for the flexible water model, namely q-TIP4P/F. The experimental free energy of water has been measured equal to -54.9 kJ/mol.^{143, 217} The difference between the computed and the experimental free energy values has to be ascribed to the presence in the real system of quantum effects that are not taken into account in the classical calculations. The energy terms are reported in Table 4.5.

Table 4.5. Hydration free energy terms computed on 300 water molecules^a

Solvent	$F_H(V)$	ΔF	FE	FE_{comp}^{146}
Water	-32.14	-2.16	-34.30	-37.62

^a All energy terms are expressed in kJ/mol.

The last term to include in the calculation of the hydration free energy is the contribution of the pure solute at standard conditions. In this context, specific computational schemes were optimized accounting for the peculiar experimental aggregation state of each solute at room temperature and atmospheric pressure.

In our application, we dealt with the liquid, solid, and vapor states.

Nitromethane, benzene, propionic acid, and piperidine are liquids at standard conditions.²¹⁸ Three independent systems including 64 molecules of nitromethane, benzene, propionic acid, and piperidine were prepared. The quasi-harmonic (QH) approximation was applied to optimize the systems' volume. The QH free energy was computed as the minimum of the QH free energy at 300 K as a function of the volume interpolated by the Birch-Murnaghan equation of state. Including the QH contribution to the system free energy, the resulting HFE estimates result to be in good agreement with the corresponding experimental values. To improve the accuracy of the QH free energy estimate for pure liquids, terms resulting from thermodynamic integration (TI) are included (Table 4.8a).

Then, we account for ketoprofen, which is solid at standard condition (Table 4.8c). In this case, the experimental crystal structure resolved by Briard and Rossi²¹⁹ and deposited in the Cambridge Crystallographic Data Centre (Accession code: CCDC 1194973) was used to compute the free energy contribution of the pure solid. An in-depth analysis of the solid state of ketoprofen is provided in Section 4.4.3.1.

The computational scheme applied to liquid and solid samples cannot be used to determine the free energy of compounds in the vapor and gas phases, such as methane and isobutane²¹⁸ because the quench of the sample would result in a population of disconnected clusters, representing the nuclei of condensation frozen by the sudden quench. As a result, the harmonic spectrum would contain a large number of zero frequency modes and the thermodynamic integration part would represent the major portion of the free energy. Therefore, it would be difficult to estimate the perturbation theory alone. On the other hand, the low density of vapor and

gas samples makes them suitable to be investigated by Monte Carlo in its grand-canonical variant (GC-MC, Sec. 2.1.4 and 2.5).

In our implementation, we start from a small number of molecules ($v \sim 12$) in a volume large enough to contain ~ 100 molecules at the experimental vapor equilibrium density. The system is simulated and μ is changed while tuning μ in order to achieve the experimental average density. At the end of this procedure, one obtains the chemical potential $\Delta\mu$ to be added to the intra-molecular free energy, f_{intra} , including the vibrational f_{vib} , the rotational f_{rot} , and ideal f_{id} free energy contributions, to obtain the free energy of a single molecule in the equilibrium phase at standard conditions, providing the last ingredient to compute the solvation free energy of pure solutes in the vapor and gas phases. At this stage, we identify μ with the excess Helmholtz free energy, although in principle one needs to integrate $\partial f / \partial N = \mu$.

Table 4.6 lists the vibrational free energies, f_{vib} , for all the compounds included in the series. In our application, only the f_{vib} values including the QH and perturbation contributions for the single solute molecule of methane and isobutane are used.

Table 4.6. Vibrational free energies for the series of globally neutral small molecules^a

Solutes	$F_H(V)$	ΔF	f_{vib}
Methane	50.55	0.52	51.07
Isobutane	160.98	0.58	161.56
Nitromethane	28.48	-0.53	27.95
Benzene	115.36	0.1	115.46
Propionic acid	81.90	-21.47	60.43
Piperidine	187.72	0.10	187.82
Ketoprofen	272.39	1.21	273.60

^a All energy terms refer to the single solute molecule and are expressed in kJ/mol.

Rotational free energies, f_{rot} , have been computed in the classical limit, because at 300 K the spacing of rotational energies are already well below the thermal energy.

In the classical limit, we first compute the inertia tensor of the molecule in its ground state:

$I_{xx} = \sum_{ia=1}^{na} m_{ia}(y_{ia}y_{ia} + z_{ia}z_{ia})$	$I_{xy} = - \sum_{ia=1}^{na} m_{ia}x_{ia}y_{ia}$	$I_{xz} = - \sum_{ia=1}^{na} m_{ia}x_{ia}z_{ia}$	(4.41)
$I_{xy} = - \sum_{ia=1}^{na} m_{ia}x_{ia}y_{ia}$	$I_{yy} = \sum_{ia=1}^{na} m_{ia}(x_{ia}x_{ia} + z_{ia}z_{ia})$	$I_{yz} = - \sum_{ia=1}^{na} m_{ia}y_{ia}z_{ia}$	
$I_{zx} = - \sum_{ia=1}^{na} m_{ia}x_{ia}z_{ia}$	$I_{zy} = - \sum_{ia=1}^{na} m_{ia}y_{ia}z_{ia}$	$I_{zz} = \sum_{ia=1}^{na} m_{ia}(x_{ia}x_{ia} + y_{ia}y_{ia})$	

This small 3×3 symmetric matrix is diagonalized to obtain the three principal momenta of inertia, Ω_x , Ω_y , and Ω_z .

As a second step, rotational *temperatures*, Θ_x , Θ_y , and Θ_z are computed:

$\Theta_x = \frac{\hbar^2}{2\Omega_x k_B}$	$\Theta_y = \frac{\hbar^2}{2\Omega_y k_B}$	$\Theta_z = \frac{\hbar^2}{2\Omega_z k_B}$	(4.42)
--	--	--	--------

where k_B is the Boltzmann factor that is in the denominator to transform an energy into a temperature.

As a final step, the partition function is evaluated and defined as follows:

$Q_R = \frac{\pi^{1/2}}{\sigma} \left(\frac{T^3}{\Theta_x \Theta_y \Theta_z} \right)^{1/2} = e^{-\beta F_R}$	(4.43)
---	--------

where σ is the number of symmetry operations for the ground state molecule ($\sigma = 12$ for methane, $\sigma = 12$ for benzene, etc...).

Table 4.7 lists the rotational free energies, f_{rot} , for all the compounds included in the series. The numerical results show that the contribution of f_{rot} on the HFEs is sizeable. The values refer to the gas-phase molecule and, as such, they concern directly the reference free energy of low density methane and isobutane. In solution, rotations are partially hindered and the rotational entropy is modified. The rotational contribution to the system free energy cancel out for solutes that are in the liquid state at standard conditions, whereas it is included considering solutes that are in the solid state at normal conditions (i.e. ketoprofen).

Table 4.7. Rotational free energies at 300 K for the series of globally neutral organic molecules^a

Solutes	f_{rot}
Methane	-7.07
Isobutane	-25.24
Nitromethane	-23.91
Benzene	-20.25
Propionic acid	-26.32
Piperidine	-27.35
Ketoprofen	-36.76

^a Energy terms are expressed in kJ/mol.

Table 4.8a-c summarizes our HFE estimates for the selection of globally neutral small molecules. In particular, Table 4.8a lists the compounds that are in the liquid state at standard conditions; Table 4.8b includes methane and isobutane that are vapors at standard conditions; Table 4.8c reports ketoprofen, which is solid at standard conditions.

In the following paragraph, we specified how the terms reported in Table 4.8a-c have been computed.

The perturbation term, ΔF , for the hydrated solutes results from the difference of the contributions regarding the hydrated solute and bulk water, considering the system of 300 water molecules:

	$\Delta F = [\Delta F_{A,wat} - \Delta F_{wat}]_{300\ mol}$	(4.44)
--	---	--------

where $[\Delta F_{wat}]_{300\ mol}$ is equal to -647.88 kJ/mol (see Table 4.5).

For solutes that are in the liquid and vapor states at standard conditions (Tables 4.8a and 4.8b), the free energy difference $(FE_{A,wat} - FE_{wat})$ results from:

	$(FE_{A,wat} - FE_{wat}) = [F_H(V)_{A,wat} - F_H(V)_{wat}]_{300\ mol} + \Delta F$	(4.45)
--	---	--------

where $[F_H(V)_{wat}]_{300\ mol}$ is the QH free energy of 300 SPC/Fw water molecules and is equal to -9642.77 kJ/mol (see Table 4.5). Note that in this case the ideal free energy contributions of the pure solute and the system including the hydrated solute cancel out in the difference defining the HFE.

For ketoprofen (Table 4.8c), the rotational free energy is added to the free energy terms referring to the hydrated solute, because its contribution to the HFE is not included in the terms computed for the pure solute in the solid state. Finally, the difference between the ideal free energy contributions (Δf_{id}) computed on the system of ketoprofen in water and on the solid sample is included because these terms do not cancel out in the difference defining the HFE.

	$(FE_{A,wat} - FE_{wat}) = [F_H(V)_{A,wat} - F_H(V)_{wat}]_{300\text{ mol}} + \Delta F + f_{rot} + \Delta f_{id}$	(4.46)
--	---	--------

Two main strategies have been adopted to determine the free energy term regarding the pure solute, FE_A , depending on the aggregation state of the solute at standard conditions. Note that the FE_A value reported in Table 4.8a-c refers to one solute molecule.

For solutes listed in Table 4.8a, which are in the liquid state at standard conditions, and for ketoprofen (Table 4.8c) that is solid at normal conditions, FE_A is obtained as:

	$FE_A = F_H(V)_A + \Delta F_A$	(4.47)
--	--------------------------------	--------

where $F_H(V)_A$ is the QH free energy and ΔF_A is the perturbation contribution.

For methane and isobutane (Table 4.8b), which are vapors at standard conditions, FE_A is computed as:

	$FE_A = f_{vib} + f_{id} + F_{GC-MC} + f_{rot}$	(4.48)
--	---	--------

where f_{vib} is the vibrational free energy (see Table 4.6), f_{id} is the ideal contribution, F_{GC-MC} is the free energy term computed by GC-MC, and f_{rot} is the rotational free energy of the single solute molecule. The intra-molecular, f_{intra} , includes the vibrational, ideal, and rotational free energy contributions (f_{vib} , f_{id} , and f_{rot} , respectively)

Finally, the hydration free energy, HFE , is obtained as:

	$HFE = (FE_{A,wat} - FE_{wat}) - FE_A$	(4.49)
--	--	--------

Table 4.8a-c. HFE estimates for the series of globally neutral small molecules^a

- a. Free energy terms of compounds that are in the liquid state at standard conditions. Note that one should add the ideal free energy contributions depending on the volume concentrations of each molecular species. These terms nearly cancel out in the difference defining the HFE. For this reason, they have not been added in Table 4.8a.

$$[F_H(V)_{\text{wat}}]_{300 \text{ mol}} = -9642.72 \text{ kJ/mol}, [\Delta F_{\text{wat}}]_{300 \text{ mol}} = -647.88 \text{ kJ/mol}$$

Solutes	Hydrated solute			Pure solute			HFE	HFE _{exp}	HFE _{FreeSolv}
	$F_H(V)_{A,\text{wat}}$	ΔF	$FE_{A,\text{wat}} - FE_{\text{wat}}$	$F_H(V)_A$	ΔF_A	FE_A			
Nitromethane	-9782.17	86.17	-53.29	-34.22	-2.57	-36.79	-16.50	-16.74	-8.70
Benzene	-9603.33	8.26	47.64	54.16	0.18	54.34	-6.70	-3.64	-3.39
Propionic acid	-9689.17	30.50	-15.96	3.69	-0.99	2.70	-18.65	-27.03	-38.03
Piperidine	-9613.30	85.52	114.93	111.11	-0.83	110.28	4.65	-21.42	-16.19

- b. Free energy terms of compounds that are in the vapor state at standard conditions.

Solutes	Hydrated solute			Pure solute					HFE	HFE _{exp}	HFE _{FreeSolv}
	$F_H(V)_{A,\text{wat}}$	ΔF	$FE_{A,\text{wat}} - FE_{\text{wat}}$	f_{vib}	f_{id}	$F_{\text{GC-MC}}$	f_{rot}	FE_A			
Methane	-9689.95	68.26	21.03	51.07	-29.00	-5.60	-7.07	9.40	11.63	8.24	10.25
Isobutane	-9631.68	84.36	95.40	161.56	-41.51	-10.40	-25.24	84.41	10.99	9.67	10.63

- c. Free energy terms of ketoprofen, which is solid at standard conditions

Solute	Hydrated solute				$FE_{A,\text{wat}} - FE_{\text{wat}}$	Pure solute			HFE	HFE _{exp}	HFE _{FreeSolv}
	$F_H(V)_{A,\text{wat}}$	ΔF	f_{rot}	Δf_{id}		$F_H(V)_A$	ΔF_A	FE_A			
Keto	-9525.78	11.13	-36.76	-8.93	82.37	106.18	-12.05	94.13	-11.76	-45.10	-72.13

^a The QH free energies of the hydrated solutes, $F_H(V)_{A,\text{wat}}$, are multiplied by 300 (i.e. the number of water molecules in the system). The QH free energies of the pure solutes, $F_H(V)_A$, refer to one molecule. The experimental HFE_{exp} values are those summarized by Martins et al.²⁰² The computed HFE_{comp} values refer to the FreeSolv database.¹⁹⁹ All energy terms are expressed in kJ/mol.

From our computations, the HFE estimates computed for the neutral form of propionic acid, piperidine, and ketoprofen do not agree with the experimental data. However, in reality, the hydrated forms of propionic acid and ketoprofen are dissociated ($HA + H_2O \rightleftharpoons A^- + H_3O^+$), whereas piperidine is protonated ($B + H_2O \rightleftharpoons BH^+ + OH^-$). The presence of a charged solute affects the hydrogen-bonding network of bulk water and its entropy giving a non-negligible contribution to the system HFE. In order to include this effect, an analytical correction of the computed HFEs was applied considering the relationship between the Gibbs free energy difference and the dissociation constants, pK_a (Eq. 4.50). For propionic acid and piperidine, the IUPAC experimental values of the dissociation constants were considered (i.e. 4.87 and 11.28, respectively).²²⁰ For ketoprofen, the experimental pK_a reported in the Drug Bank Database was used (i.e. 4.45).²²¹⁻²²²

	$\Delta G = 2.303RTpK_a$	(4.50)
--	--------------------------	--------

Table 4.9 summarizes the HFE estimates of the series including the correction factor for propionic acid, piperidine, and ketoprofen.

Table 4.9. HFE estimates for the series of globally neutral small molecules including the correction factor accounting for the protonation state of propionic acid, piperidine, and ketoprofen^a

Solutes	Hydrated solute	Pure solute	pK _a correction			
	$FE_{A,wat} - FE_{wat}$	FE_A	ΔG	HFE	HFE _{exp}	HFE _{FreeSolv}
Methane	21.03	9.40	-	11.63	8.24	10.25
Isobutane	95.40	84.41	-	10.99	9.67	10.63
Nitromethane	-53.29	-36.79	-	-16.50	-16.74	-8.70
Benzene	47.64	54.34	-	-6.70	-3.64	-3.39
Propionic acid	-15.96	2.70	27.98	-46.63	-27.03	-38.03
Piperidine	114.93	110.28	15.62	-10.97	-21.42	-16.19
Ketoprofen	82.37	94.13	25.56	-37.32	-45.10	-72.13

^a The contribution due to the protonation state was added to the HFE with the negative sign accounting for the increased stability in water of the charged species. All energy terms are expressed in kJ/mol.

In the following Section 4.4.3.1, an in-depth analysis of the vibrational properties ketoprofen in the solid state is reported.

4.4.3.1. Detailed analysis of ketoprofen in the equilibrium crystal phase

Ketoprofen, a propionic acid derivative, is the largest solute molecule considered in this study and it is also the one with the most apparent pharmacological interest, being a nonsteroidal anti-inflammatory agent (NSAIA) with analgesic and antipyretic properties. Because of its large size, and since it is the only solute which is in a crystal at standard conditions, the checks we performed on the force field and on the general properties of this molecule are discussed.

A major ingredient in the computation of the solvation energy of ketoprofen is the force field used to model its PES, whose reliability is assessed here by comparison of computed and measured properties of ketoprofen in the crystal phase.

In Figure 4.23, Left, the experimental unit cell of solid ketoprofen is reported. The torsional angle of $\sim 25^\circ$ between the two aromatic rings (due to the short contacts of the carbonyl oxygen with hydrogens in the rings) prevents the molecule to stay in the planar conformation. Once in the crystallographic cell, the two molecules of ketoprofen adopt an anti-parallel orientation characterized by the aromatic cycles facing each other oriented at $\sim 71^\circ$. Replicating the unit cell on the x-axis (Fig. 4.23, Right), two hydrogen bonds are established between two molecules located in adjacent unit cells giving insight into the chemical and structural reasons of the stability of the crystal structure as well as of the low solubility of neutral ketoprofen.

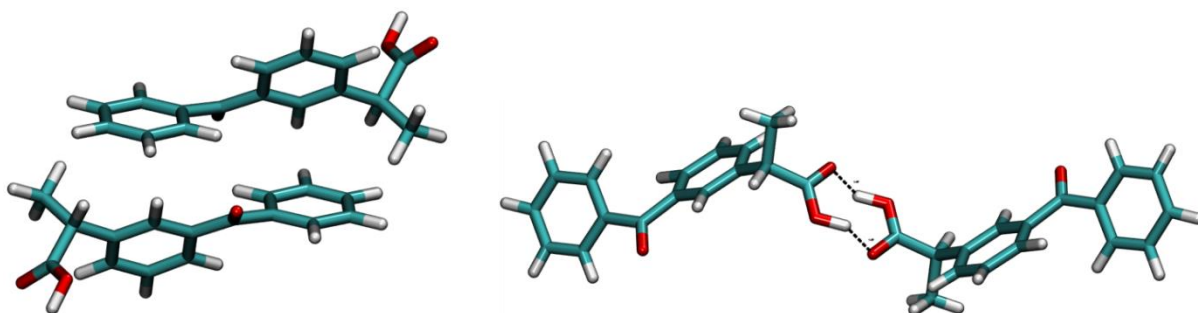


Figure 4.23. (Left) Crystallographic cell including two molecules of ketoprofen. (Right) Hydrogen bonds established between two ketoprofen molecules of adjacent cells.

As a required step to benchmark the force field in the calculation of the structure and of vibrational frequencies of ketoprofen, density functional (DFT) computations have been carried out for ketoprofen in the experimental crystal structure.

Standard DF theory in the Kohn-Sham formulation⁶⁴ has been used, with a generalized gradient approximation for the exchange-correlation potential.⁶⁶ Dispersion interaction, missing in the original GGA-PBE approximation, have been reintroduced using the empirical prescription by Grimme.²²³ This recipe, although approximate and somewhat empirical, is known to provide a qualitatively correct description of the structure and vibrational properties of the molecular crystals, including those whose structure is affected by hydrogen bonding.

In our computations, Kohn-Sham orbitals have been expanded in plane waves, with a kinetic cut-off of 120 Ry. Only valence electrons are included in the computation, and their interaction with the underlying closed-shell core is described through norm-conserving ab initio pseudo-potentials.²²⁴

The starting geometry is provided by the atomistic lattice parameters measured by X-ray diffraction.²¹⁹ The simulated sample consists of a single unit of two ketoprofen molecules and 66 atoms (38 non-hydrogen atoms) of triclinic symmetry, $P_{\bar{1}}$ (Fig. 4.23, Left). The lattice parameters are listed in Table 4.10. Because of the relatively large unit cell and insulating character of this molecular crystal, in the electronic structure computation the sampling of the Brillouin zone has been limited to the Γ -point only.

Table 4.10. Lattice parameters of ketoprofen^a

a	b	c	α	β	γ	V
13.893	7.741	6.136	89.61	94.56	88.78	657.639

^a Cell lengths a , b , and c , are expressed in Å; angles α , β , and γ in degrees; cell volume, V , in Å³.

First, the geometry has been optimized to minimize the energy by quenched MD, consisting of constant energy MD stretches terminated by quenches whenever $\sum_i \mathbf{v}_i \cdot \mathbf{f}_i$ becomes negative. The lattice parameters have been kept at the experimental value. The optimal structure is very close to the experimental one, having a square deviation per atom, χ^2 , defined as:

	$\chi^2 = \sum_{i=1}^n \frac{ \mathbf{r}_i^{(out)} - \mathbf{r}_i^{(in)} ^2}{n}$	(4.51)
--	--	--------

equal to 0.0387 Å² when the sum runs on the non-hydrogen atoms, and 0.0665 Å² when the sum covers all atoms. As expected, the deviation is higher for hydrogen atoms, whose position is rather uncertain in X-ray diffraction measurements. To be precise, the structure reported in the experimental paper²¹⁹ had one H atom less than the stoichiometric definition, which has been added to our input by standard software tools.

The tighter identification of H atoms by DFT is reflected in the better definition of the pair of H-bonds at the carboxylic junction between two molecules, which appears stronger and more planar in the computational ground state structure than in the experimental one.

The carbonyl group at the center of the molecules does not accept a proper hydrogen bond, but nevertheless it forms weak hydrogen bonds with hydrogens belonging to aromatic rings.

Vibrational frequencies have been computed by diagonalising the dynamical matrix (i.e. the Hessian matrix weighted by the atomic masses) computed by finite difference. Also in this case, only the Γ -point of the Brillouin zone has been considered.

The analysis of the eigenvalues and eigenvectors shows that the vibrational density of states (vDOS) extends from $\omega \sim 0$ to $\omega = 3150 \text{ cm}^{-1}$ (Fig. 4.24).

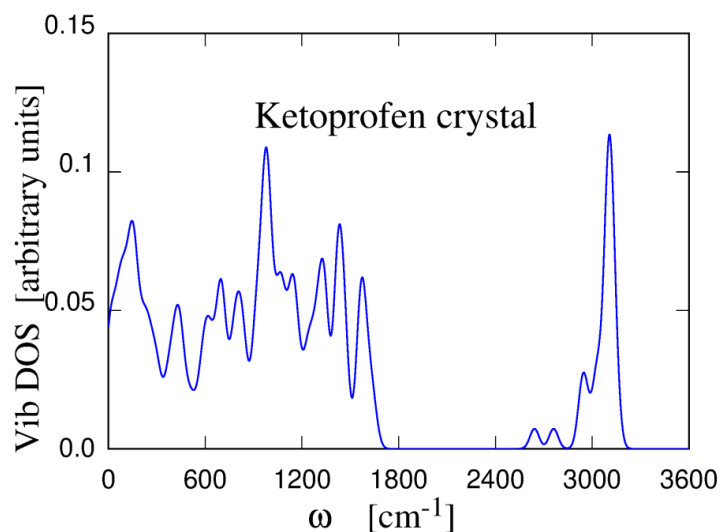


Figure 4.24. Vibrational density of states of crystal ketoprofen from DFT calculations.

The high frequencies band from $\omega = 3090 \text{ cm}^{-1}$ to $\omega = 3150 \text{ cm}^{-1}$ consist of $\text{C}(\text{sp}^2)\text{-H}$ stretching modes, while $\text{C}(\text{sp}^3)\text{-H}$ stretching modes occur from $\omega = 2940 \text{ cm}^{-1}$ to $\omega = 3040 \text{ cm}^{-1}$ (Fig. 4.25).

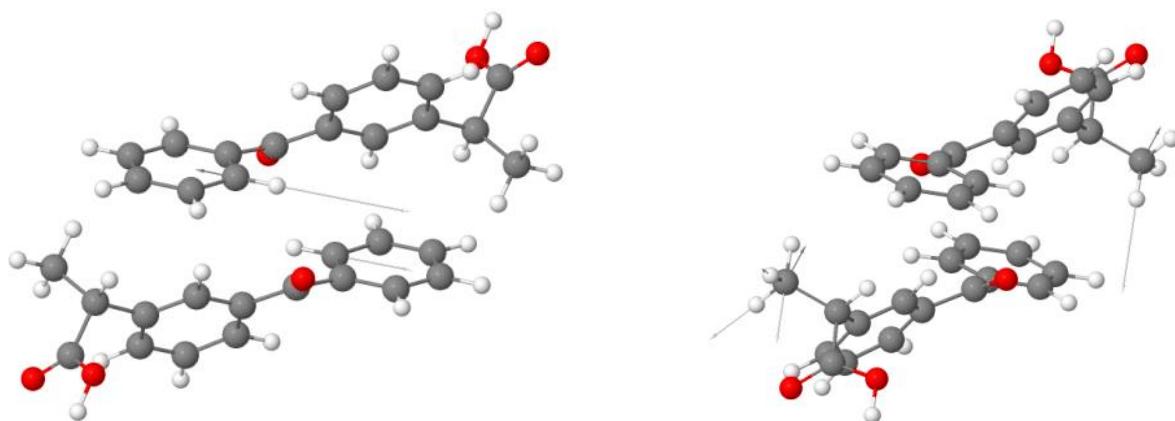


Figure 4.25. (Left) $\text{C}(\text{sp}^2)\text{-H}$ stretching modes at $\omega = 3144 \text{ cm}^{-1}$. (Right) $(\text{sp}^3)\text{-H}$ stretching modes at $\omega = 3040 \text{ cm}^{-1}$.

The only $\text{C}(\text{sp}^3)\text{-H}$ bond (per molecule) belonging to CH_3 groups, gives origin to a very narrow doublet (two bonds per unit cell) at $\omega = 2955 \text{ cm}^{-1}$ (Fig. 4.26).

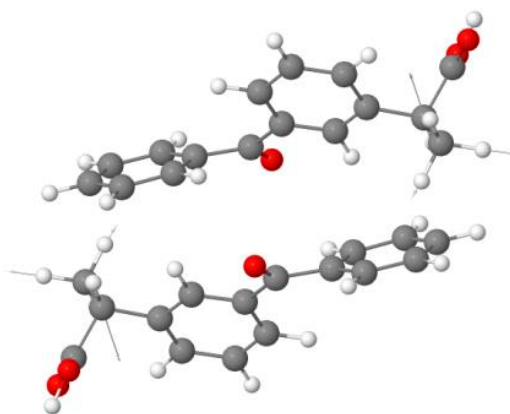


Figure 4.26. C(sp³)-H stretching modes belonging to CH₃ groups at $\omega = 2955 \text{ cm}^{-1}$.

The two O-H stretching modes (per unit cell) occur at $\omega = 2640 \text{ cm}^{-1}$ and $\omega = 2760 \text{ cm}^{-1}$, their frequency being split and lowered by the formation of inter-molecular hydrogen bonds (Fig. 4.27).

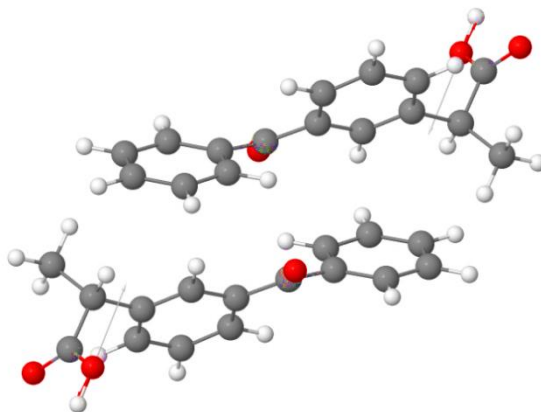


Figure 4.27. O-H stretching modes at $\omega = 2640 \text{ cm}^{-1}$.

The stretching of the aromatic C=C and C-O bonds form a band covering $1550\text{-}1650 \text{ cm}^{-1}$. The C-O bonds, in particular, tend to occupy the top of the band (Fig. 4.28).

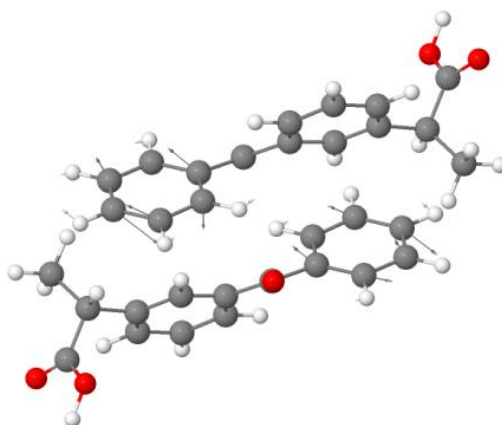


Figure 4.28. Stretching modes of the aromatic C=C and C-O bonds at $\omega = 1557 \text{ cm}^{-1}$.

C=C-H and C-C-H bending modes are found at $1050\text{-}1470 \text{ cm}^{-1}$. Below 1000 cm^{-1} , the vibrational eigenvalues show that modes are rather hybridized, mixing bending of different bond types and torsional modes below $\sim 600 \text{ cm}^{-1}$ (Fig. 4.29).

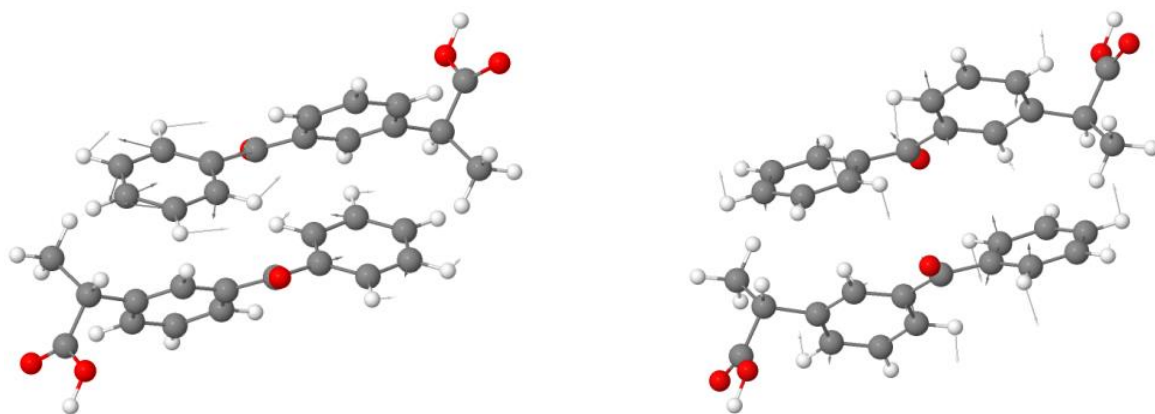


Figure 4.29. (Left) C=C-H and C-C-H bending modes at $\omega = 1468 \text{ cm}^{-1}$. (Right) Low frequency modes at $\omega = 920 \text{ cm}^{-1}$.

Although no experimental data are available on the vibrational properties of crystalline ketoprofen, the DFT results are consistent with what is known from experiments on systems made of organic molecules of similar size and structure. For this reason, we use the DFT data to benchmark the force field results.

The force field QH free energy of ketoprofen in the solid state was computed replicating $2 \times 3 \times 3$ times the experimental crystallographic cell, giving a simulated sample of 36 molecules and 1188 atoms.

Since the angle between the experimental lattice vectors is nearly 90° , the structure of crystal ketoprofen has been optimized in a slightly deformed supercell of orthorhombic shape and lattice vectors $a = 27.752$, $b = 23.196$, and $c = 18.387 \text{ \AA}$. In this case, the comparison of the QH free energies resulting from the classical and quantum calculations is carried out at equal volume of the unit cell.

Once again, the system volume was optimized minimizing the classical harmonic free energy for each temperature with respect to the volume. In this case, the interpolation of the volume dependence of the free energy at 300 K was smoother in comparison to the system of water, because of the limited number of configurations that ketoprofen can adopt in the crystal phase (Fig. 4.30).

The resulting minimum volume, $V_0 = 337 \text{ \AA}^3$, per molecule is 2.2% larger than the experimental $V_0 = 329 \text{ \AA}^3$ per molecule.

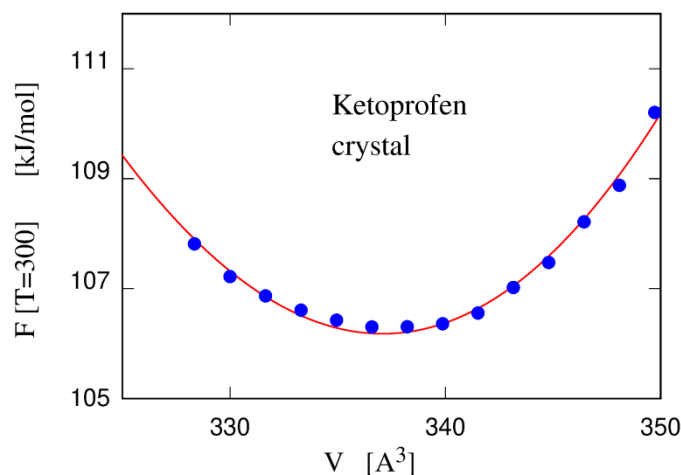


Figure 4.30. Interpolation of the volume dependence of the free energy at 300 K computed for ketoprofen at the solid state.

For each volume tested for the QH approximation, the reciprocal orientation of the ketoprofen molecules in the unit cell can be compared, providing additional information on the stability of the crystal changing the volume. In our application, a slightly distorted conformation was observed when the system volume was increased by 6% with respect to the experimental one.

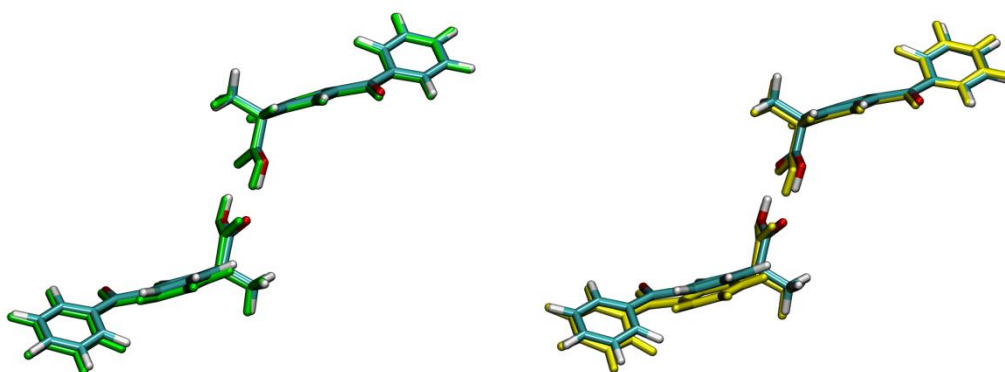


Figure 4.31. Superimposition of ketoprofen molecules in crystal phase changing the system volume. Comparison of the experimental reciprocal orientation in the crystallographic unit cell (cyan colored) with the orientation obtained increasing the system volume by 4.5% (left, green colored) and 6% (right, yellow colored).

Comparison of the ground state structure shows that the experimental structure is fairly well reproduced also by the force field, having a mean square deviation (per atom), χ^2 , equal to 0.11 \AA^2 when considering non-hydrogen atoms, and 0.47 \AA^2 when considering all atoms. Visual inspection of the structure shows that the nearly perfect planarity of the phenyl rings displayed the experimental structure is reproduced by the force field. As expected, discrepancies are observed in the dihedral angles, whose definition is often somewhat uncertain when using the force field.

The good agreement of structural properties is reflected into the fair agreement of the vibrational density of states (vDOS). The comparison of the vDOS computed by DFT and by the force field is shown in Figure 4.32.

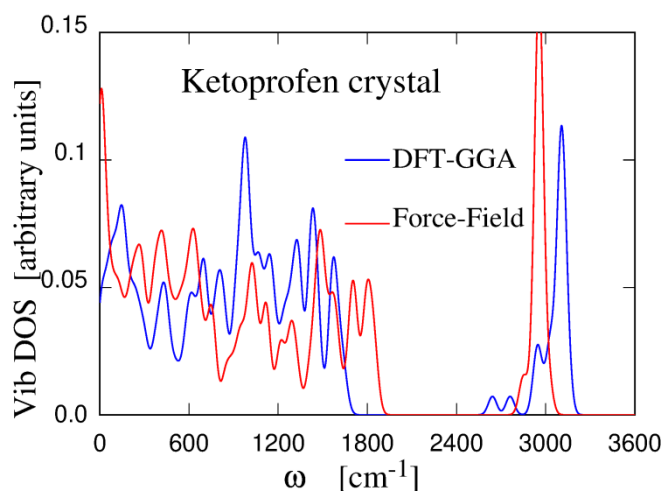


Figure 4.32. Comparison between the vDOS computed by DFT (blue) and by the force field (red).

Despite quantitative discrepancies, it is possible to identify the same stretching, bending, torsion, and molecule-molecule displacements in both vDOS. The major difference is a low frequency peak that is apparent in the force field result and absent in the DFT vDOS.

The similarity of the two vDOS underlies the good agreement of the vibrational contribution to the free energy per molecule shown in Figure 4.33.

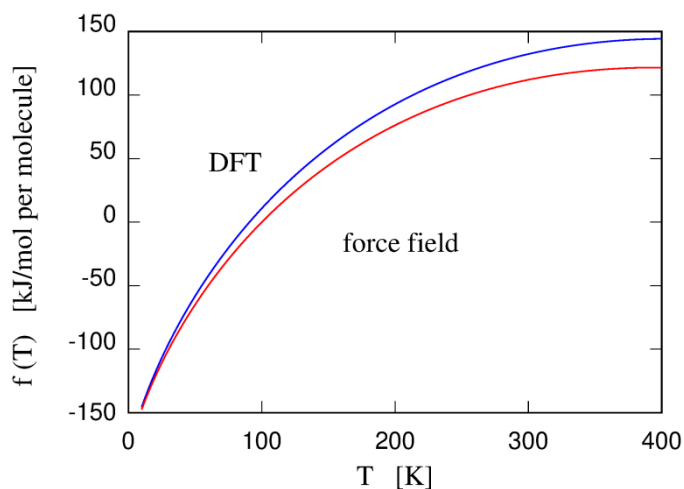


Figure 4.33. QH free energy per molecule of ketoprofen at the solid state as a function of temperature. Comparison between the results computed by DFT (blue) and by the force field (red).

The force field could be further fine-tuned to provide an optimal fit of the DFT vibrational free energy upgrading the quality of the force field simulation to a near DFT level.

Notice that, in comparison to the other compounds of the series, the free energy of the hydrated solute of ketoprofen includes the ideal and the rotational contributions. The former is added as the difference between the ideal free energy terms computed on the system of ketoprofen in water and on the solid sample in order to compensate the fact that these terms do not cancel out in the difference defining the HFE. The latter has to

be included in the HFE estimate of every solute in the solid state at normal conditions, including ketoprofen, being the rotations of the solute prevented in the solid sample.

By including a large solute in our series, such as ketoprofen, we realized that the rotational free energy, severely limited by the (many) residual harmonic restraints even at $\lambda = 0.995$, have to be taken into account, and, if necessary, manually added to the HFE estimate computed with our approach. Moreover the ideal free energy term does not cancel out in the comparison of the three systems relevant for the solvation free energy computation since the pure solvent ideal free energy is reduced by the fact that ketoprofen is a crystal at standard conditions. This term, however, is easily accounted for without any further approximation.

We here emphasize that the rotational free energy contribution should not be included explicitly by applying the alternative approach presented at the end of Section 4.4.2.1 based on performing the perturbation dynamics at low temperature located below the glass temperature of water. At this low energy state, the rotational free energy contribution is small. Moreover, it is completely recovered during the perturbation dynamics increasing the system temperature of the fully-interacting system up to 300 K.

4.5. Scaling to large systems

As already mentioned, by definition, the Hessian is a $(3N \times 3N)$ matrix, whose elements are the second derivatives of the potential energy with respect to the coordinates of the $3N$ atoms. As such, it can be interpreted in terms of harmonic springs connecting pairs of atoms. Computing and diagonalizing the Hessian matrix as a dense matrix might be computationally time consuming for large systems.¹⁹⁶ However, despite the long range of Coulomb forces, the Hessian matrix is primarily a sparse matrix as soon as the system size exceeds a few hundred molecules. This is because most of the Coulomb energy is already summarized by the potential energy of the local minimum, while the range of electrostatic interactions arising from polar distortions of the structure are effectively screened to short range (see Perfect screening theorem²²⁵). Moreover, the intrinsic disordered of all the local minima we considered, confines the elastic interactions that otherwise could also propagate to long range.

We verified that the degree of sparsity can be pushed to the limit of a banded matrix with ~ 24 off-diagonal elements. Widely available library routines diagonalize a banded matrix of the size $(3N \times m)$ with a number of operations of the order of $m \times (3N)^2 + m^2 \times (3N)$, but more advanced algorithms such as *divide and conquer*, reduce the cost to linear scaling from the original $(3N)^3$ scaling for dense matrices. For systems of 10^5 atoms, such as those of pharmacological interest, the saving can range from a factor of 10^5 for quadratic algorithms to a (somewhat optimistic) factor of 10^{10} for linear scaling. It is apparent that in this size range sparsity and linear scaling are mandatory.

Somewhat surprisingly, the cost of filling the Hessian can easily exceed the cost of diagonalizing it, even as a dense matrix. In fact, there are $(3N)^2$ matrix elements in the Hessian, each requiring at least four energy evaluations, each taking a number of operations that scales linearly, or, more usually, quadratically with the number N of atoms.

To speed up the filling of the Hessian and to set the stage for its diagonalization as a sparse matrix, a careful analysis of the matrix structure was carried out looking at the relationship between force constants and distance among interacting atoms. In particular, the minimal tetrahedral model with hydrogen atoms located on tetrahedral vertices around each oxygen atom has been identified as a promising lead (Fig. 4.34).

Thus, small clusters of interacting atoms centered on one oxygen atom were identified in the system of 300 water molecules, applying a cut-off radius long enough to include at least the second neighboring tetrahedral clusters. Then, the analysis was repeated for clusters centered on hydrogen atoms.

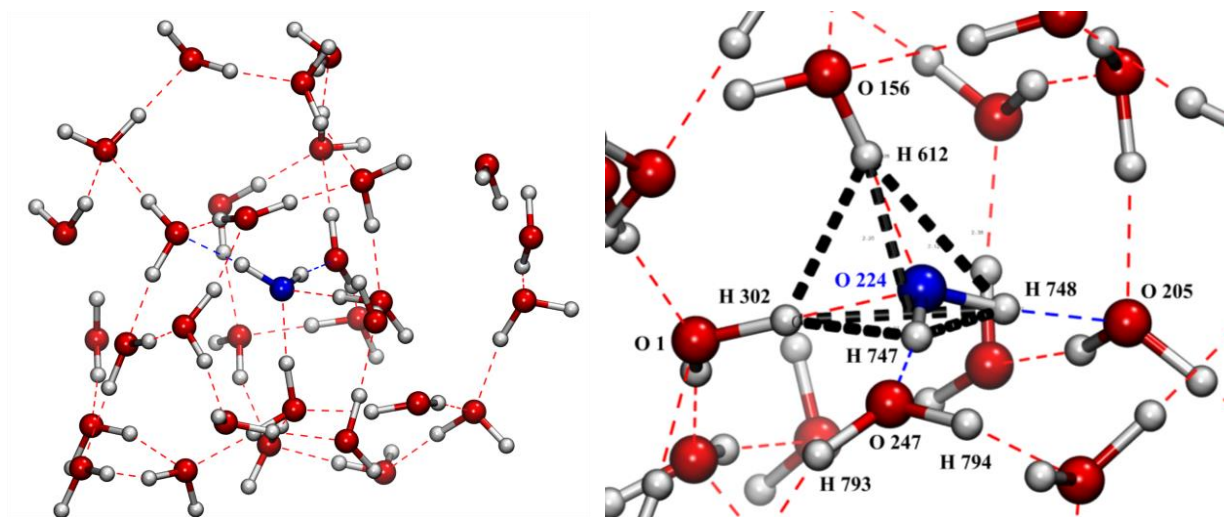


Figure 4.34. (Left) Representation of hydrogen bond network within 6 Å centered on the blue colored O224. (Right) Minimal tetrahedral cluster centered on O224. Hydrogen bonds are blue and red colored; harmonic springs connecting the hydrogen atoms on tetrahedral vertices are reported as black dashed lines. The index numbering refers first to 300 oxygens and then to 600 covalently bonded hydrogen atoms (e.g. Oa, Ob, etc., H1a, H2a; H1b, H2b, etc...).

As apparent from the definition, each atom pair i and j corresponds to a (6×6) Hessian sub-matrix defined as:

	$k = \frac{\partial^2 U}{\partial R_i^\alpha \partial R_j^\beta}$	(4.52)
--	---	--------

Neglecting the diagonal (3×3) blocks corresponding to $i = j$, and taking into account the symmetry of second derivatives, the crucial portion is one of the two off-diagonal 3×3 sub-blocks referring to the interactions of atoms $i \neq j$. Since $\mathbf{R}_I - \mathbf{R}_J$ separation has, in principle, an isotropic orientation, it is difficult to extract from the 3×3 matrix an intuitive picture of the two-body interaction. To clarify the picture, we

diagonalize the matrix, interpreting the eigenvalues as spring constants acting along the direction of the three orthogonal eigenvectors. The matrix is symmetric and the eigenvalues are real, although not necessarily positive.

In all cases we analyzed, the diagonalization gives one eigenvector that to a good approximation lies along the $i - j$ direction, and two eigenvectors that are (again approximatively) perpendicular to this direction. For relatively short $i - j$ separations the eigenvalue of the longitudinal eigenvector is significantly larger than those of the transverse eigenvectors. We interpret the first eigenvalue-eigenvector as representing a harmonic stretching spring between i and j . The other two (eigenvalue-eigenvector) pairs are interpreted in term of shear elastic springs.

In what follows, we focus on the stretching component, under the assumption that tangential forces are accounted for by the network of springs surrounding any given atom pair. The stretching contribution, in particular, was analyzed based on the definition of harmonic force constants, which are dependent on the atomic coordinates only. In particular, the analysis refers to the minimal tetrahedral model represented in Figure 4.34, Left. The atom numbers are those in input to our MD simulations.

In Table 4.11, the stretching distances and force constants between O224 and neighboring atoms within a distance of 3.2 Å are reported.

Table 4.11. Stretching distances and force constants between O224 and neighboring atoms within a distance of 3.2 Å^a

ia	ja	dr	$\log(dr)$	k	$\log(-k)$
224	748	1.037E+00	3.600E-02	-4.473E+03	8.406E+00
224	747	1.044E+00	4.320E-02	-4.558E+03	8.425E+00
224	612	1.603E+00	4.716E-01	-1.132E+02	4.729E+00
224	302	1.632E+00	4.897E-01	-1.072E+02	4.675E+00
224	156	2.641E+00	9.712E-01	-5.672E+02	6.341E+00
224	247	2.649E+00	9.742E-01	-5.465E+02	6.304E+00
224	1	2.666E+00	9.806E-01	-5.051E+02	6.225E+00
224	205	2.693E+00	9.905E-01	-4.475E+02	6.104E+00
224	709	2.955E+00	1.083E+00	-1.790E+01	2.885E+00
224	415	2.956E+00	1.084E+00	-1.787E+01	2.883E+00

224	786	2.999E+00	1.098E+00	-1.709E+01	2.839E+00
224	611	3.050E+00	1.115E+00	-1.624E+01	2.787E+00
224	301	3.195E+00	1.161E+00	-1.411E+01	2.647E+00

^a Distances, dr , and force constants, k , are expressed in Å and kJ/Å², respectively.

The first coordination shell is properly defined. The covalently bonded hydrogen atoms (i.e. 747, 748) are located at 1.04 Å from the central O 224, and the hydrogen-bonded H atoms (i.e. 302, 612) were identified at 1.60 Å and 1.63 Å. Within 2.70 Å, the four O atoms included in the first tetrahedral coordination shell were identified (i.e. 1, 156, 205, and 247).

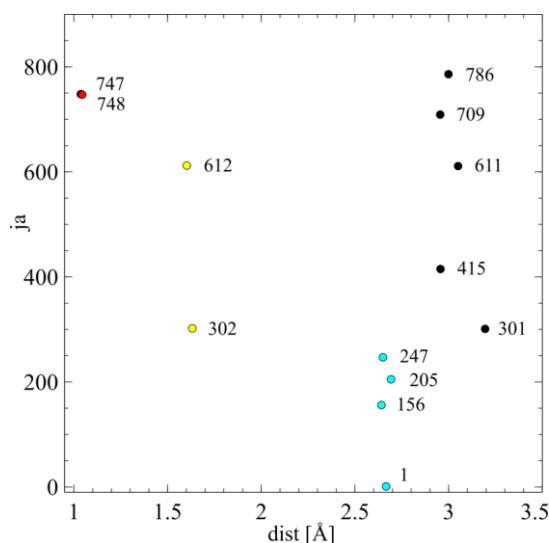


Figure 4.35. Stretching distance as a function of the atoms index of atoms in the system of 300 water molecules. The list of interacting atoms was cut at 3.2 Å. Covalently bonded H atoms ($dr = 1.0$ Å) are red colored. Hydrogen-bonded H atoms ($dr = 1.6$ Å) and four O atoms included in the first tetrahedral coordination shell ($dr = 2.6$ - 2.7 Å) are colored in yellow and cyan, respectively. Atoms located at greater distances are colored in black.

Then, force constants associated with harmonic springs connecting O224 and each neighboring atom are evaluated.

In Figure 4.36, the stretching distance between the central O224 and neighboring atoms as a function of the force constants in logarithmic scale is reported. As expected, harmonic springs connecting O224 and the covalently bonded hydrogen atoms are associated to force constants close to the experimental OH stretching (i.e. $\sim 4,500$ kJ/Å²). Then, two clearly separated force constants are identified. They refer to hydrogen-bonded H atoms (i.e. 302, 612) and four O atoms included in the first tetrahedral coordination shell (i.e. 1, 156, 205, and 247), respectively.

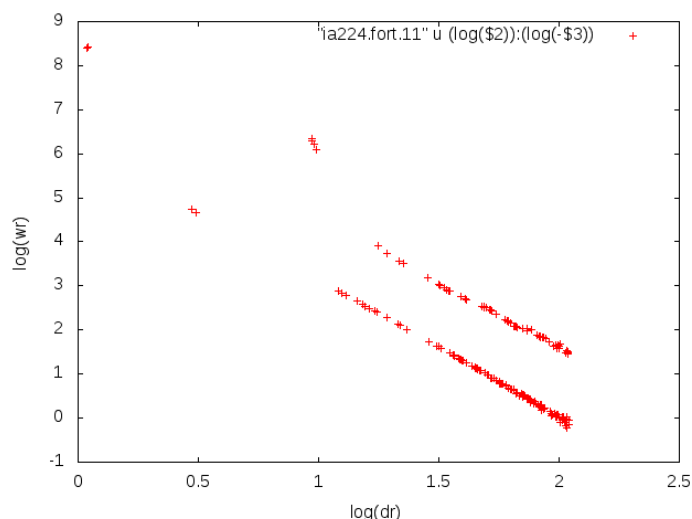


Figure 4.36. Stretching distance between the central O224 and neighboring atoms as a function of the force constants in logarithmic scale.

Interestingly, harmonic springs describing non-covalent interactions between O224 and H atoms suggested weaker hydrogen bonds if compared with the ones associated with interactions between O224 and O atoms, despite OO interactions are mediated by one H atom, by definition. In the mesh-like representation of the solvent, oxygen atoms determined the overall structure of the system due to their mass and dimension, which are significantly higher than those of hydrogen atoms located in the mesh interstices. Therefore, higher values of OO force constants were supposed to be related to the collective nature of interacting hydrogen bonding network.

As a first guess, we assumed that harmonic springs describing covalent OH stretching and non-covalent interactions between OH and OO were related with each other according to the following relationships. Therefore, supposing the force constants k_1 as OH stretching and the k_2 as describing the OH hydrogen bond, the sum of both contributions (i.e. $k_{1,2}$) matched the harmonic spring describing OO interaction, which resulted to be slightly greater than the one characterizing OH hydrogen bond.

	$\Delta x_1 = \frac{F}{k_1}$	$\Delta x_2 = \frac{F}{k_2}$	(4.53)
	$k_{1,2} = \frac{F}{\Delta x_{1,2}} = \frac{F}{\frac{F}{k_1} + \frac{F}{k_2}} = \frac{1}{\frac{1}{k_1} + \frac{1}{k_2}} = \frac{k_1 k_2}{k_1 + k_2}$		(4.54)
	$\frac{1}{k_{1,2}} = \frac{1}{k_1} + \frac{1}{k_2}$		(4.55)

To verify the consistence of the force constant values changing the atom identity of the cluster's centers, we improved the description of the system centering the cluster on one of the two covalently bonded H atoms to oxygen atom 224 (i.e. H747).

In Table 4.12, the stretching distances and the force constants between characterizing the network centered on H747 within a distance of 3.3 Å are reported.

Table 4.12. Stretching distances and force constants between H747 and neighboring atoms within a distance of 3.3 Å^a

ia	ja	<i>dr</i>	log(<i>dr</i>)	<i>k</i>	log(− <i>k</i>)
747	224	1.044E+00	4.320E-02	-4.321E+03	8.371E+00
747	247	1.611E+00	5.321E-01	-1.115E+02	5.517E+00
747	748	1.702E+00	4.768E-01	-2.489E+02	4.714E+00
747	302	2.074E+00	7.294E-01	-5.204E+01	3.952E+00
747	612	2.119E+00	7.510E-01	-4.877E+01	3.887E+00
747	794	2.248E+00	8.100E-01	-4.085E+01	3.710E+00
747	793	2.445E+00	8.941E-01	-3.201E+01	3.466E+00
747	1	2.979E+00	1.092E+00	-1.744E+01	2.859E+00
747	156	3.041E+00	1.112E+00	-1.635E+01	2.794E+00
747	611	3.272E+00	1.185E+00	-1.328E+01	2.587E+00
747	205	3.290E+00	1.191E+00	-1.288E+01	2.555E+00

^a Distances, *dr*, and force constants, *k*, are expressed in Å and kJ/Å², respectively.

Evaluating distances between H747 and neighboring atoms as a function of atoms indexes, the covalently bonded oxygen atom was properly identified at 1.04 Å (i.e. 224). The second covalently bonded H atom to O 224 (i.e. H748) and the O atom hydrogen-bonded to H747 (i.e. O247) were located at 1.70 Å and 1.61 Å, respectively. The hydrogen-bonded H atoms to O224 were identified at 2.07 Å and 2.12 Å, respectively (i.e. 302, 612). Two H atoms covalently bonded to O247 were identified at 2.25 Å and 2.44 Å (i.e. 793, 794).

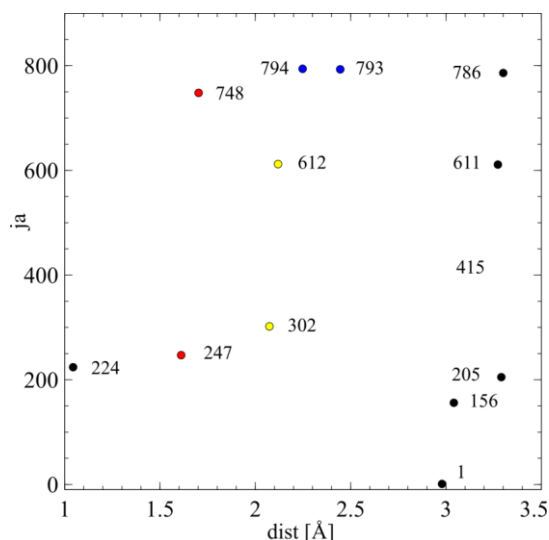


Figure 4.37. Stretching distance between H747 and neighboring atoms as a function of the particles indexes for the system of 300 water molecules. The list of interacting atoms was cut at 3.3 Å. Covalently bonded O224 ($d_r = 1.0$ Å) is colored in black. The second covalently bonded H atom to O224 and the O atom hydrogen-bonded to H747 (H748 and O247, $d_r = 1.6$ - 1.7 Å) are red colored. The hydrogen-bonded H atoms to O224 (H302 and H612, $d_r = 2.1$ Å) are colored in yellow. Two H atoms covalently bonded to O247 (H793 and H794, $d_r = 2.2 - 2.4$ Å) are reported in blue.

Evaluating the corresponding force constants, the interaction between H747 and the covalently bonded O224 was characterized by a force constant value close to OH stretching (i.e. $-4,320$ kJ/Å²). In the mesh-like description of liquid water, the introduction of a harmonic spring between the covalently bonded H atoms to O224 (i.e. 747, 748) defined an implicit bending term. Interestingly, the entity of the harmonic spring connecting covalently bonded H atoms to O224, H747, and H748, was comparable to the one describing the interaction between H747 and the adjacent hydrogen-bonded O247 (i.e. -248.9 and -111.5 kJ/Å², respectively), giving substantial information about how neighboring clusters were connected with each other.

Therefore, distortions of the selected minimal tetrahedral cluster were evaluated, looking at harmonic springs connecting H747 and hydrogen-bonded H atoms to O224 (i.e. 302, 612). In the present case, those interactions were associated with similar force constants (i.e. 52.0 and 48.8 kJ/Å²), defining a symmetric tetrahedron. It was observed that interactions between H747 and H atoms covalently bonded to the adjacent O247 (i.e. 794, 793) were associated with force constants ranging from -40.9 and -32.0 kJ/Å². Interactions between H747 and the O atoms included in the second coordination shell, apart from O247, were associated with weak harmonic springs, ranging from -17.4 and -12.9 kJ/Å².

In Figure 4.38, the stretching distance between the central H747 and neighboring atoms as a function of the force constants in logarithmic scale is reported.

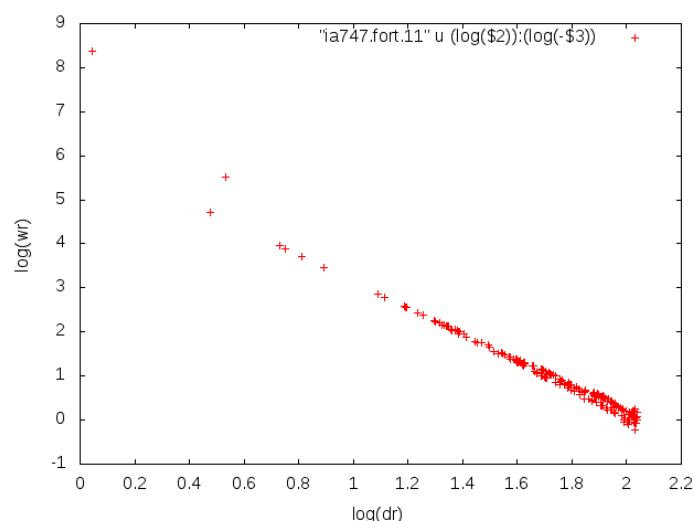


Figure 4.38. Stretching distance between the central H747 and neighboring atoms as a function of the force constants in logarithmic scale.

To completely evaluate the geometry of the selected minimal tetrahedral cluster, harmonic springs connecting H atoms hydrogen-bonded to O224 (i.e. 302, 612), and H atoms 302 and 612 with the second covalently bonded H atom of O224 (i.e. 748) were considered.

Table 4.13. Stretching distances and force constants among H302, H612, and H748^a

ia	ja	dr	$\log(dr)$	k	$\log(-k)$
302	612	2.196E+00	7.866E-01	-4.380E+01	3.780E+00
302	748	2.384E+00	8.689E-01	-3.441E+01	3.538E+00
612	748	2.381E+00	8.674E-01	-3.455E+01	3.542E+00

^a Distances, dr , and force constants, k , are expressed in Å and kJ/Å², respectively.

H302 and H612 were located at 2.20 Å, and they were connected by harmonic spring characterized by a force constant equal to -43.80 kJ/Å². Harmonic springs connecting H748 with H302 and H612, were characterized by distances of 2.38 Å and force constants equal to -34.41 and -34.55 kJ/Å², respectively, contributing to cluster's symmetry.

To improve the comprehension of the network established by hydrogen bonds in liquid water, minimal tetrahedral clusters centered on 6 different oxygen atoms were compared evaluating the force constants as a function of distance. The minimal tetrahedral cluster centered on O247 has a structural defect, due to the presence of only three oxygen atoms in the first coordination shell instead of 4 (Fig. 4.39). The presence of defective tetrahedrons has to be evaluated carefully, limiting the proper reconstruction of the hydrogen bonding network of the overall system.

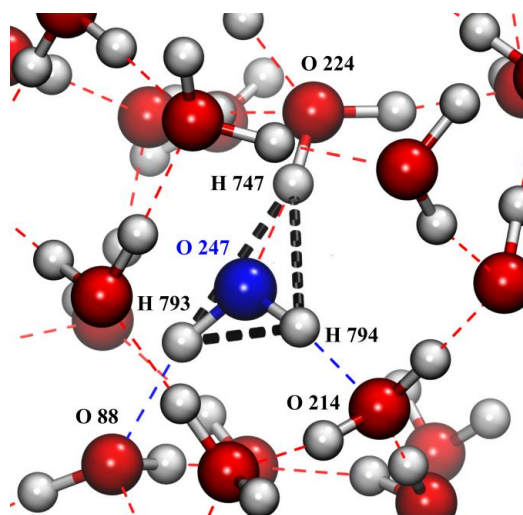


Figure 4.39. Minimal tetrahedral cluster centered on O247. Hydrogen bonds were reported in blue and red. Harmonic springs connecting hydrogen atoms included in the first coordination shell were black colored forming a planar triangular hydrogen bonding network.

Nevertheless, the distribution of the force constants with respect to inter-atomic distances suggests a regular hydrogen-bonding network centered on O247.

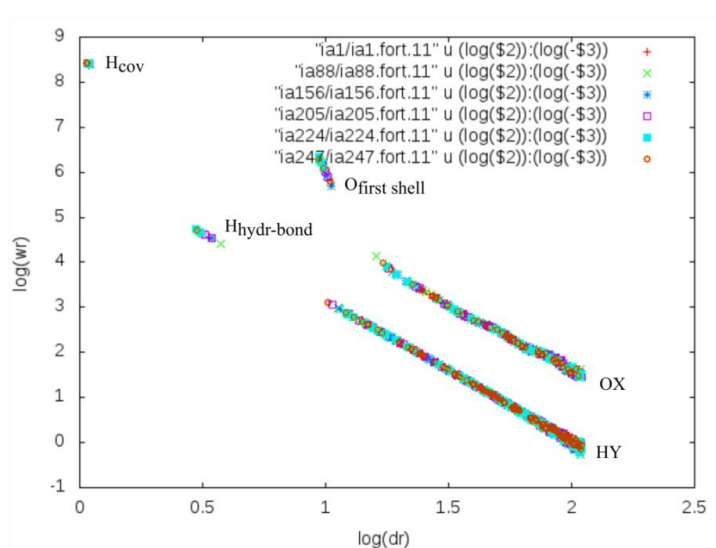


Figure 4.40. Superimposition of force constants as a function of distance in logarithmic scale obtained by 6 different minimal tetrahedral clusters centered on O atoms.

From the comparison of all clusters, three groups of force constants referring to atoms included in the first coordination shell, (i.e. covalently bonded H atoms, hydrogen-bonded H atoms, and four O atoms), are clearly identified.

In conclusion, from this analysis emerged a predictable logarithmic trend where force constants associated to atoms included in the first coordination shell are clearly identifiable. Therefore, in principle, the values of the force constants referring to the atoms included in the first coordination shell can be used to compute the Hessian as a sparse matrix, without losing information regarding the hydrogen-bonding network of the overall water system. To speed up the filling of the Hessian matrix, the linearly fitted force constants related

to the harmonic springs connecting minimal tetrahedral clusters can also be considered, exploiting the linear trend of the force constant with respect to the inter-atomic distances in logarithmic scale. Following this approach based on the reconstruction of the Hessian matrix, the computation of the Hessian scales linearly with the dimension of the system.

To validate this assumption, we filled the Hessian matrix of 300 water molecules including the short-range stretching contributions only. From the diagonalization of the Hessian, the vibrational density of state (vDOS) of the system modeled by harmonic springs was provided (Fig. 4.41).

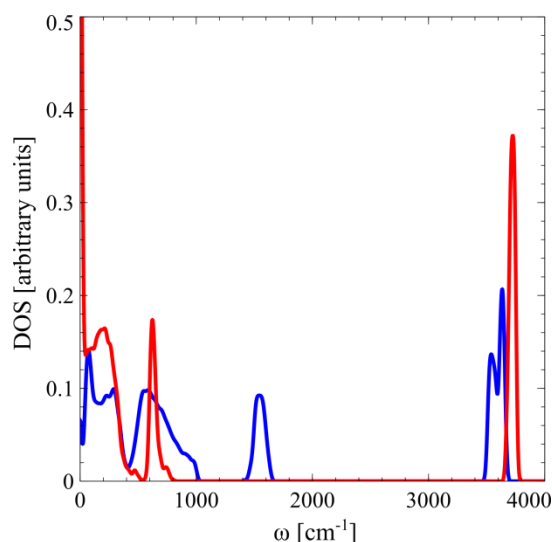


Figure 4.41. Comparison between vDOS of the system of 300 water molecules resulting from the diagonalization of the Hessian matrix filled with longitudinal stretching contributions (blue) and of the full Hessian matrix (red).

High frequency OH stretching and bending can be recognized, despite the corresponding frequencies are over- and under-estimated, respectively. Low frequency modes result to be affected by the low number of inter-molecular interactions that were taken into account. Thus, further evaluations need to be considered to use this approach to speed up the computation of the Hessian matrix reproducing accurately the hydrogen-bonding network of a water system.

Another approach to improve the efficiency of the computation of the Hessian matrix was explored.

The idea is to fill the Hessian including the minimum number of interacting particles whose minima were located within a cut-off distance, resulting in the correct vDOS of water. To minimize the effects of the limited number of inter-atomic interactions on the low frequency modes, the cut-off distance has to be carefully optimized based on the desired speed-accuracy trade-off.

The cut-off distance was defined looking at the list of neighboring particles around each oxygen atom and identifying the corresponding covalently bonded hydrogen atoms.

In Table 4.14, cut-off distances and number of particles included in the neighboring list are reported.

Table 4.14. Cut-off distances and corresponding number of interacting atoms computed on the test system of 300 water molecules^a

$dr_{\text{cut-off}}$	N. atoms
16.6	899
3.6	28
3.2	21
2.2	7
1.6	3

^a Cut-off distances, $dr_{\text{cut-off}}$, are expressed in Å

As expected, the cut-off equal to 1.6 Å resulted in a distorted vibrational density of state (vDOS). The cut-off set at 2.2 Å including only atoms that are part of the minimal tetrahedral model, seemed to correctly represent the vibrational states of water. The cut-off distance equal to 3.6 Å was considered as a good compromise between speed of calculation and accuracy in the representation of the system vibrations (Fig. 4.42).

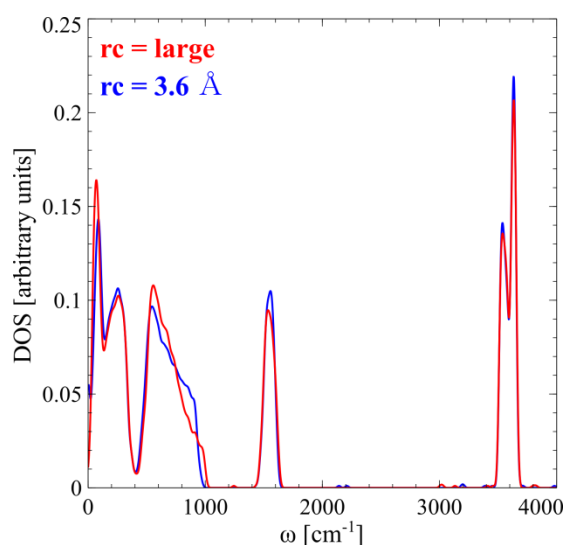


Figure 4.42. Superimposition of the vDOS obtained considering the entire hydrogen-bonding network (red) and setting the cut-off distances at 3.6 Å (blue)

An additional ingredient useful to make this approach applicable to large systems would be the reduction of the number of simulations required to recover the full an-harmonicity of the system. In this context, the integrand might be symmetrized with respect to $\lambda = 0.5$ and then fitted by:

	$\frac{1}{x + \delta} \times \frac{1}{ 1 - x + \delta}$	(4.56)
--	--	--------

This type of approach has been tested with promising but preliminary results.

To summarize the observations discussed in this Section 4.5, the analysis of the full Hessian matrix computed for one of the bulk water samples including 300 SPC/Fw water molecules that we generated, suggests that the Hessian matrix can be reduced to a banded matrix by computing the Hessian elements only for atoms whose distance is less than a pre-set cut-off.

Moreover, the computation of the matrix elements from the force field might be replaced by the much cheaper approach of (accurately) estimating matrix elements through a simple model of harmonic interactions among neighboring atoms.

The task of reducing the cost of computing the Hessian is greatly eased by the relative freedom in choosing the reference Hamiltonian (provided it is harmonic), since the perturbation step will nevertheless fill the gap between the reference and actual systems.

4.6. Discussion and conclusions

In this chapter of the thesis, a method to compute free energies applicable to systems of arbitrary complexity has been introduced. Our computational scheme relies on computing the absolute free energy including the quasi-harmonic term arising from the vibrational modes of the system, the ideal contribution, and the an-harmonic free energy computed by thermodynamic integration.

We started from the application of the confinement method to biological systems proposed by Karplus. Here, the reference system relies on the Einstein approximation, from which a simple analytical expression for the free energy involving a single frequency can be derived.

First, we improved the definition of the reference system replacing the Einstein model with the Debye model, which requires a more complex analysis but provides a better representation of the real system. In the framework of the Debye approximation, all the vibrational modes are obtained by the diagonalization of the Hessian matrix computed on the atomic coordinates of an arbitrary local energy minimum. The eigenvalues computed by the diagonalization of the Hessian matrix give access to the system harmonic free energy. To reduce the dependence of the harmonic free energy estimate on the local minimum, we applied the quasi-harmonic approximation, which allows us to obtain the volume-dependent free energy fitting the harmonic free energy values changing the system volume by an analytic equation of state.

The system an-harmonicity is then recovered by thermodynamic integration. In this step of our protocol, the diffusion of water molecules emerges as a limitation in the computation of the integrand of thermodynamic integration.

We proposed some approaches to overcome this problem. One of them is computing free energy differences, such as hydration free energy (HFE), whose estimate requires the consideration of three independent samples: hydrated solute, bulk water, and pure solute at normal conditions. In the difference of energy terms (including the perturbation contribution that is integrated up to $\lambda = 0.995$), the effect of diffusion of water molecules was supposed to cancel out subtracting the bulk water contribution to the term related to the solvated system. By computing HFEs of generic solutes, the aggregation state at standard conditions needs to be available in order to provide accurate free energy estimates.

To enhance the correlation among the local minima found by the three independent samples, each system is briefly annealed lowering the temperature from 300 K to 150 K, and then quenched to the nearest local minimum. This step minimizes the difference of the quasi-harmonic free energies between the hydrated solute system and bulk water.

To test our approach and implementation, we computed the classical and quantum absolute free energy of a flexible water model, in addition to the HFEs of a selection of globally neutral small molecules including one

of pharmacological interest (i.e. ketoprofen). The results are in very good agreement with the experimental data. Regarding ketoprofen, which is the largest solute molecule and the only one in the solid state at standard conditions, we underlined the importance of explicitly include the rotational free energy contribution in our HFE estimate. The agreement of experimental and computational solvation energies is remarkable, because of the intrinsic difficulty in estimating this quantity, and, even more, because of the crucial role played by the solvation free energy in a wide variety of biochemical processes.

For the modest system sizes considered in our study, the TI accounts for 80% of the computational time. The computational effort required for the diagonalization increases with the system size. In Section 4.5, possible approaches to improve the operations of filling and diagonalizing the Hessian matrix have been discussed as factors limiting the application of the method to large systems of pharmacological interest.

Validating and testing our computational scheme, we realized that the accuracy of the HFE estimates crucially depend on a number of hidden parameters that the user has to define depending on the system under investigation. Making the computational scheme independent from details regarding the system and the protocol itself, is required to make our approach widely applicable to systems including thousands of atoms.

We also realized that the free energy of the pure solute at normal conditions gives a sizeable contribution to the solvation free energy, which is a point that remains not properly discussed in the literature regarding previous applications of related methods to compute absolute free energies.

On the other hand, the choice of the reference Hamiltonian still leave much freedom that could be exploited to reduce the weight of the TI step. The restrained centers do not need to be positions of minimum energy and also the force constants do not need to be derived from the diagonalization of the Hessian matrix.

In conclusion, our approach based on free energy differences integrated up to $\lambda = 0.995$ is able to provide accurate HFE estimates of globally neutral small molecules solvated in explicit water. Some challenges need to be addressed to improve the accuracy of the estimates and the computational efficiency.

In order to overcome the problem of water diffusion resulting in the integrand divergence for $\lambda \rightarrow 1$, we also presented other approaches that rely on limiting the harmonic Hamiltonian to a finite value. To this aim, we first make the harmonic restraint periodic in space. As such, when a water molecule moves closer to a periodic replica of the minimum energy position than the original one, we adopt the periodic replica as the origin of the restraint. Despite this approach seems promising, the force of the restraint suddenly changes during the simulation, making it unstable. Then, we implemented and tested a couple of previously published approaches based on swapping the water molecules to minimize their harmonic free energy, as described in the literature. Our conclusions suggest that these simulation approaches are problematic in terms of the high computational cost and the rigorous identification of the simulation ensemble, making the results unreliable and unusable to estimate free energies for systems in explicit solvent.

We also explored an alternative approach to limit the role of water diffusion in the $\lambda \rightarrow 1$ limit by performing the perturbation dynamics from the harmonic model to the real system at low temperature, located below the glass temperature of water. The quasi-harmonic term is then added to the perturbation contribution to obtain the absolute free energy of the system at low temperature. Finally, the free energy of the fully-interacting system at 300 K is recovered by including the free energy gained by the system increasing the temperature up to 300 K. The preliminary results are very promising although further calculations are needed to improve statistics.

5. Conclusions

In this thesis, I described how molecular dynamics (MD)-based methods can be applied to address two open challenges in computational medicinal chemistry: unbinding kinetic predictions and free energy estimates.

The former has received increasing attention in recent years, since Copeland has proposed the residence time (i.e. the reciprocal of the dissociation rate constant) as a good estimator of the *in vivo* efficacy of biologically active compounds. The latter is a challenging task since decades. Indeed, accurate free energy estimates can be obtained only by extensively sampling the complex energy landscapes characterizing biomolecular systems, which is not straightforward. Although nowadays the increasing computer power makes it possible the dynamical explorations of complete protein-ligand binding processes, the accurate characterization of the thermodynamic landscape is still challenging.

After a general overview of the thermodynamic and kinetic aspects of inter-molecular interactions, I provided the theoretical framework relating both thermodynamics and kinetics to free energy, which is in fact the *fil rouge* of the overall discussion presented in this thesis.

In this context, I worked on a new computational methodology to simulate protein-ligand dissociation events that has been presented in Chapter 3. As introduced in Chapter 2, transitions between free energy basins separated by barriers larger than few $k_B T$ represent *rare events* relative to the time scales accessible through plain MD simulations. This is because of the exponential relationship between the rate at which the barrier is crossed and the free energy describing the system kinetics and thermodynamics, as stated by transition state theory (TST). As a result, the high energy configurations (i.e. transition states) are much less frequently visited than low-energy states. To overcome the limitations due to the Boltzmann sampling, we developed an *enhanced* MD-based protocol combining the adiabatic-bias molecular dynamics (ABMD) with an electrostatics-like collective variable, dubbed eLABMD, to facilitate the dissociation of protein-ligand binary complexes.

ABMD is based on the definition of a representative reaction coordinate, to which a time-dependent harmonic biasing potential is added. In particular, by specifying the extremal points to be joined, it is possible to monitor whether or not the system evolves spontaneously toward the target state. Therefore, the pawl-and-ratchet biasing potential is added to the potential energy function only when the system attempts to move in the opposite direction with respect to the desired end point. This aspect makes ABMD particularly interesting giving a realistic description of the evolution of the system to an external perturbation, leaving its short-time dynamics relatively unperturbed. Moreover, by choosing the electrostatic potential between interacting entities as collective variable, we improved the description of the natural forces driving the dissociation mechanisms making our approach more physics based.

We applied the approach to two pharmaceutically relevant kinases: Glucokinase (GK) and Glycogen Synthase Kinase 3 beta (GSK-3 β). By directly leveraging the information due to the (biased) unbinding time, we were able to rank two series of ligands on unbinding kinetics. To provide statistically robust estimates of dissociation times, a modest number of independent simulations need to be collected. Additionally, we got mechanistic and path information on unbinding events. In summary, we proposed a computationally efficient methodology, which is able to provide information on relative unbinding times, together with a qualitative description of the unbinding processes. To further improve the methodology, we propose a protocol to correct the biased unbinding times obtaining the absolute, physical residence times.

The complete interpretation of equilibrium properties and kinetic transformation requires free energy considerations. Then, I worked on the implementation and validation of a new computational method to estimate absolute free energies applicable to systems of arbitrary complexity. As discussed in Chapter 2, the fundamental ingredient necessary to absolute free energy estimates is a representative reference system of known free energy. Common reference systems are the harmonic Einstein crystal for solids and the Lennard-Jones fluid for liquids.

Previous attempts to compute absolute free energy of biological systems in implicit solvent used the harmonic Einstein crystal as reference for proteins. We assumed the Debye model to be a representative reference system of both solute and solvent. In this framework, all the vibrational modes are obtained by the diagonalization of the Hessian matrix of the fully interacting system described by the force field. By the application of the quasi-harmonic approximation, we further improved the accuracy of the free energy estimate for the reference state of the system. Then, perturbation theory was applied to include the anharmonicity of the real fully-interacting system. At this stage, we explored the application of the method to solutes in explicit solvent.

This project was particularly challenging and we are still working on overcoming some issues that arose in the test applications of the method. At the present stage, however, results are already promising, and further development aims primarily to upscale the approach to large systems.

Firstly, dealing with explicit solvent, the diffusion of water molecules and its contribution to the system free energy have to be taken into account. In our implementation of the method, water diffusion causes the divergence of the integrand (but not of the integral) of thermodynamic integration. To overcome this problem, three approaches were implemented and tested: swapping the equilibrium positions throughout the simulation; changing the linear interpolation between reference and real systems into a non-linear one; and performing the perturbation dynamics to convert the harmonic reference to the real system at a temperature below the glass temperature of water, accounting for the effect of higher T by standard thermodynamic integration over a path. Although the first two approaches solve the problem of the integrand divergence, they have some limitations that have been discussed. By contrast, the preliminary results obtained by the

applying the third approach to one representative compound of the series are very promising and we are further exploring this option.

As a first application of the method, we computed the hydration free energies of a small series of globally neutral solutes. Hydration free energies represent a particular case of free energy difference, which allow us to circumvent with the divergence of the integrand of thermodynamic integration. Indeed, in the difference the effect of diffusion of water molecules was supposed to cancel out by subtracting the bulk water contribution to the term related to the hydrated solute. Therefore, the integration was limited to values of the perturbation parameter for which water molecules explore the conformational space around the local minimum without diffusing (i.e. $\lambda = 0.995$).

In the computation of hydration free energies, the estimate of the term regarding the pure solute at standard conditions has to be considered carefully. Indeed, the liquid, solid, and vapor aggregation states require different approaches to accurately compute the corresponding absolute free energy. The underlying computational details have been described and discussed.

Among the available computational methods to compute free energy, the proposed method is particularly promising providing accurate free energy estimates. However, the overall computational machinery, and in particular the computation and fill of the Hessian matrix, has to be optimized to make it applicable to large biological systems composed by thousands of atoms. In this context, we suggested possible approaches to exploit the sparsity of the Hessian matrix.

6. Acknowledgments

Achievements like the Ph.D. are always the result of years of dedication and work and it would have not been the same without the support of several people. For this reason, I want to thank who participated and supported me during this path.

First, I would like to thank my supervisor, Prof. Andrea Cavalli, for giving me the chance to join his research group in IIT. During the three years that I spent in such an inter-disciplinary working environment, I grew as professional and scientist, extremely improving my knowledge in physics and biophysics.

I am immensely grateful to Prof. Pietro Ballone, who kindly hosted me in UCD for six months, for the great help and the huge effort you invested in completing and improving my Ph.D. project. Thank you for all the scientific discussions we had and for all the things that I learnt during our meeting and talks. Your motivation to complete the work we did together mainly when things looked hard to complete was really motivating.

I want to acknowledge Dr. Sergio Decherchi, who helped me starting my Ph.D. project. Thanks for your invaluable support through this long path and for your constant presence and help. I learnt how important it is to address scientific challenges by looking carefully at each step of the process.

I would like to thank my colleagues for the pleasant working atmosphere and for the nice time we spent together.

Most importantly, I would like to express my gratitude to my family and friends, for their love, support, and constant encouragement.

7. Appendix

7.1. Median unbinding time based-ranking correlations of GK series

From the statistic of GKAs series collected setting $K = 2.0\text{E-}15 \text{ [kJ mol}^{-1}\text{]}^{-3}$, we computed the bootstrapped estimations of median eLABMD unbinding times, observing a good correlation with experimental off-rates (Spearman coefficient = 0.64). In Table A1 and Fig. A1, the results are reported.

Table A1. Experimental residence time ($t_{r,\text{exp}}$), scaled MD ($t_{r,\text{mean}}$), and eLABMD unbinding time ($t_{r,\text{median}}$) for each compound of the GK series^a

Cpd	<u>Experimental</u>		<u>Scaled MD</u>		<u>eLABMD</u>	
	$t_{r,\text{exp}}$	Rank	$t_{r,\text{mean}}$	Rank	$t_{r,\text{median}}$	Rank
		$t_{r,\text{exp}}$		$t_{r,\text{mean}}$		$t_{r,\text{median}}$
1	8.3	1	105.1 ± 10.1	1	6.63 ± 1.32	1
2	2.3	4	29.3 ± 5.3	5	5.21 ± 2.06	5
3	2.7	3	38.9 ± 7.1	4	5.77 ± 2.67	4
4	1.6	5	92.9 ± 7.3	3	9.05 ± 1.42	2
5	6.3	2	99.7 ± 6.7	2	5.86 ± 1.24	3
6 _{a,b}	0.7	6	25.9 ± 3.9	6	4.80 ± 2.07	6
7 _{a,b}	0.2	7	24.7 ± 3.0	7	2.31 ± 0.60	7

^a Experimental residence times, $t_{r,\text{exp}} = 1/k_{\text{off}}$, are expressed in s; sMD and eLABMD unbinding times ($t_{r,\text{mean}}$, $t_{r,\text{median}}$) are expressed in ns. Spearman coefficients for sMD and median eLABMD unbinding time-based ranking correlations are 0.89 and 0.64, respectively.

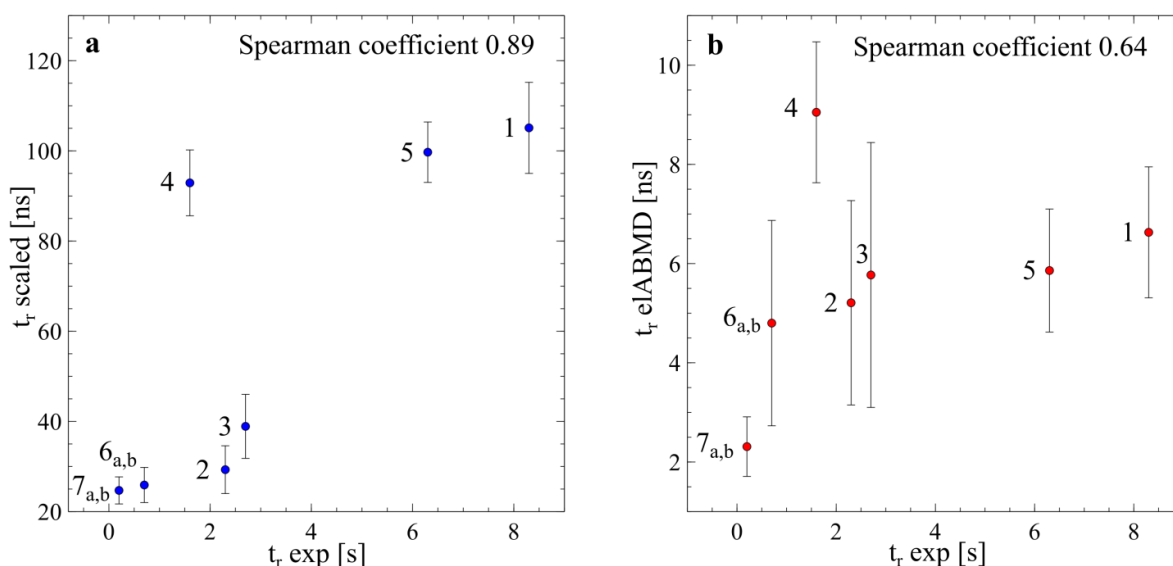


Figure A1. Experimental *versus* computational residence time. The rankings obtained with sMD (a) and bootstrapped estimations of median (b) elABMD unbinding time are reported.

High resolution bootstrapped estimations of median elABMD unbinding times (i.e., $K = 1.0E-15$ [kJ mol⁻¹]³), increased the differentiation of compounds that were poorly prioritized from a residence time standpoints at higher perturbation levels (i.e., $K = 2.0E-15$ [kJ mol⁻¹]⁻³). In Table A2, the results are reported.

Table A2. High resolution unbinding time predictions using bootstrap estimations of median unbinding times^a

<u>Exp</u>		<u>elABMD</u>			
Cpd	$t_{r,exp}$	$K = 2.0E-15$		$K = 1.0E-15$	
		$t_{r,mediann}$	rate	$t_{r,median}$	rate
5	6.3	5.86 ± 1.24	0.98	25.51 ± 6.65	0.80
3	2.7	5.77 ± 2.67		20.34 ± 8.56	

<u>Exp</u>		<u>elABMD</u>			
Cpd	$t_{r,exp}$	$K = 2.0E-15$		$K = 1.0E-15$	
		$t_{r,median}$	rate	$t_{r,median}$	rate
2	2.3	5.21 ± 2.06	0.77	23.60 ± 8.46	0.45
6 _a	0.7	4.04 ± 2.06		10.79 ± 6.69	

^a Experimental residence times ($t_{r,exp}$) are expressed in s; elABMD unbinding times ($t_{r,mean}$) are expressed in ns; the force constant, K , is expressed in [kJ mol⁻¹]⁻³. The rate between the lower and higher bootstrapped estimations of mean unbinding time was reported to highlight the increased differentiation after increasing the accuracy of unbinding simulations.

7.2. Median unbinding time based-ranking correlations of GSK-3 β series

From the statistic collected on GSK-3 β complexes, the bootstrapped estimations of median elABMD unbinding times was computed, observing a good correlation with experimental off-rates (Spearman coefficient = 0.94). In Table A3 and Fig. A2, the results are reported.

Table A3. Predicted estimations of median elABMD unbinding times ($t_{r,median}$) and experimental kinetic data for each compound of GSK-3 β series^a

Cpd	<u>elABMD</u>		<u>Experimental</u>	
	$t_{r,median}$	Rank $t_{r,median}$	$t_{r,exp}$	Rank $t_{r,exp}$
1	29.97 ± 0.26	2	609.76	1
2	29.92 ± 0.46	3	259.74	3
3	28.91 ± 1.60	5	112.04	5
4	30.00 ± 0.12	1	392.16	2
5	29.24 ± 2.20	4	219.06	4
6	17.70 ± 5.78	6	18.38	6

^a elABMD unbinding times ($t_{r,median}$) are expressed in ns and reported with an estimation of the error computed with a bootstrapped procedure. Experimental residence time are expressed in s. Spearman coefficient for median elABMD unbinding time-based ranking correlations with respect to $t_{r,exp}$ is equal to 0.94.

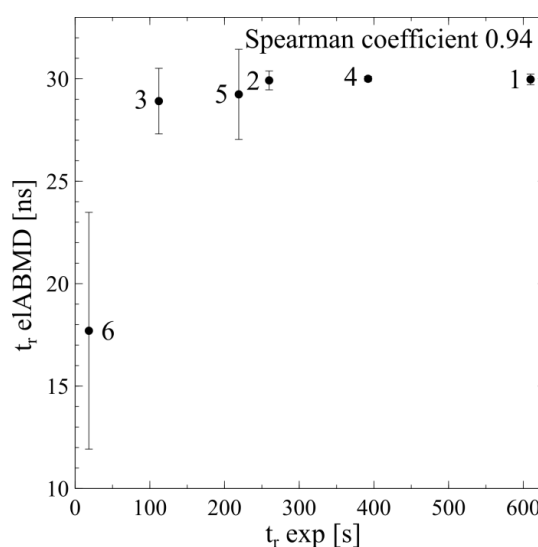


Figure A2. Experimental *versus* computational residence time. The ranking was built on the bootstrapped estimations of median elABMD unbinding times.

7.3. Statistics of randomly selected 10 production runs of GSK-3 β inhibitors unbinding simulations

To evaluate if 10 production runs were sufficient to prioritize the series of highly congeneric GSK-3 β inhibitors from a residence time standpoint, a bootstrap analysis was performed randomly selecting 10 unbinding times from those obtained from complete statistics. The procedure was repeated three times. Looking at the final results obtained from each iteration (Tables A4,A-C), a total good correlation between bootstrapped estimations of mean and median unbinding times, and experimental residence time was observed, suggesting that a statistics of 10 independent replica would be enough to perspectivevely characterize a congeneric chemical series.

Table A4,a-c. Predicted estimations of mean and median elABMD unbinding times ($t_{r,mean}$, $t_{r,median}$) computed on a randomly selected 10-replicas statistics, and experimental kinetic data for each compound of GSK-3 β series^a

- a.** Results of the first repetition of bootstrap analysis. Spearman coefficients for mean and median elABMD unbinding time-based ranking correlations with respect to experimental residence times are 0.94 and 0.60, respectively.

Cpd	<u>elABMD</u>				<u>Experimental</u>	
	Rank		Rank		Rank	
	$t_{r,mean}$	$t_{r,mean}$	$t_{r,median}$	$t_{r,median}$	$t_{r,exp}$	$t_{r,exp}$
1	28.01 \pm 1.06	1	29.60 \pm 1.11	3	609.76	1
2	25.47 \pm 2.06	4	28.16 \pm 2.48	4	259.74	3
3	25.34 \pm 2.06	5	27.92 \pm 2.43	5	112.04	5
4	27.82 \pm 1.23	2	29.63 \pm 1.09	2	392.16	2
5	27.06 \pm 2.05	3	29.82 \pm 1.10	6	219.06	4
6	16.48 \pm 2.75	6	13.89 \pm 4.93	6	18.38	6

- b.** Results of the second repetition of bootstrap analysis. Spearman coefficients for mean and median, elABMD unbinding time-based ranking correlations with respect to experimental residence times are 0.77 and 0.77, respectively.

Cpd	<u>eLABMD</u>				<u>Experimental</u>	
	$t_{r,mean}$	Rank	$t_{r,median}$	Rank	$t_{r,exp}$	Rank $t_{r,exp}$
		$t_{r,mean}$		$t_{r,median}$		
1	28.00 ± 0.94	2	29.11 ± 1.30	3	609.76	1
2	28.26 ± 1.63	1	29.99 ± 0.38	1	259.74	3
3	24.62 ± 2.12	4	27.04 ± 3.43	5	112.04	5
4	27.92 ± 1.33	3	29.82 ± 1.06	2	392.16	2
5	23.94 ± 2.50	5	27.11 ± 4.47	4	219.06	4
6	23.39 ± 2.46	6	25.85 ± 3.20	6	18.38	6

- c. Results of the third repetition of bootstrap analysis. Spearman coefficients for mean and median eLABMD unbinding time-based ranking correlations with respect to experimental residence times are 0.71 and 0.77, respectively.

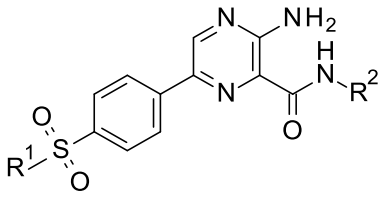
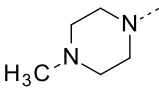
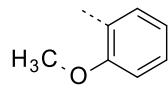
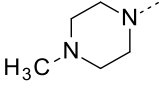
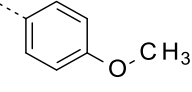
Cpd	<u>eLABMD</u>				<u>Experimental</u>	
	$t_{r,mean}$	Rank	$t_{r,median}$	Rank	$t_{r,exp}$	Rank $t_{r,exp}$
		$t_{r,mean}$		$t_{r,median}$		
1	29.73 ± 0.92	3	29.11 ± 1.30	3	609.76	1
2	29.82 ± 0.17	1	29.99 ± 0.38	1	259.74	3
3	25.08 ± 1.80	4	27.04 ± 3.43	5	112.04	5
4	29.80 ± 0.89	2	29.82 ± 1.06	2	392.16	2
5	24.50 ± 2.38	5	27.11 ± 4.47	4	219.06	4
6	19.34 ± 2.84	6	25.85 ± 3.20	6	18.38	6

^a eLABMD unbinding times ($t_{r,mean}$, $t_{r,median}$) are expressed in ns and reported with an estimation of the error computed with a bootstrapped procedure. Experimental residence times are expressed in s.

7.4. Compounds 5 and 7

Compound **7** was included in the highly congeneric chemical series of GSK-3 β inhibitors, differing by compound **5** by the position of the methoxy group to the phenyl ring (Table A5). Interestingly, despite the subtle structural modification, the potency¹⁶⁸ of those compounds differed by a factor of 57.5 thus making those ligands a challenging workbench. Compound **7** was prepared according to the general synthetic procedure showed in Scheme 3.1 and 3.2, but due to low solubility, kinetic rates of compound **7** were not determined by surface plasmon resonance (SPR).

Table A5. Chemical structure of synthesized compounds 5 and 7.

		
Cpd	R ¹	R ²
5		
7		

The methoxy oxygen atom of the substituent in ortho position to the phenyl ring of compound **5** was engaged in an intra-molecular interaction with the adjacent amide nitrogen atom, absent in compound **7**. The para position of the methoxy group in compound **7** prevented the alignment of the terminal methoxy phenyl ring to the amide group hindering Phe67 to optimally close the binding site. Therefore, fast computational unbinding time of compound **7** (Table A6) were supposed to be related to the positioning of the hydrophilic methoxy functionalities directed towards the solvent promoting the rapid solvation of the binding site.

Table A6. Predicted estimations of eLABMD unbinding times ($t_{r,mean}$, $t_{r,median}$), experimental potency (K_i) for compounds **5 and **7**, and experimental residence time of compound **5**^a**

Cpd	<u>eLABMD</u>		<u>Experimental</u>
	$t_{r,mean}$	$t_{r,median}$	$t_{r,exp}$
5	24.94 ± 1.70	29.24 ± 2.20	219.06
7	20.93 ± 1.89	21.49 ± 4.61	nd

^a eLABMD unbinding times ($t_{r,mean}$, $t_{r,median}$) are expressed in ns, K_i in nM, experimental residence time in s. eLABMD predictions were reported with an estimation of the error computed with a bootstrapped procedure.

Taking into account the key role of solvation in determining protein-ligand complex lifetime, we deeply investigated the influence of methoxy group positioning in attracting water molecules to the binding site. Supposing the para-methoxyphenyl group of compound **7** able to promote the rapid solvation of the binding site, we computed the probability maps associated to water molecules distribution around compounds **5** and **7** in their bound states employing the Hydra Analysis module included in the BiKi Life Science 1.3 software package.¹⁷⁴ As expected, a reduced water density was computed around the ortho-methoxy phenyl ring of compound **5**, whereas the para-methoxy group of compound **7** attracted a higher number of water molecules promoting the rapid solvation of the binding site (Fig. A3).

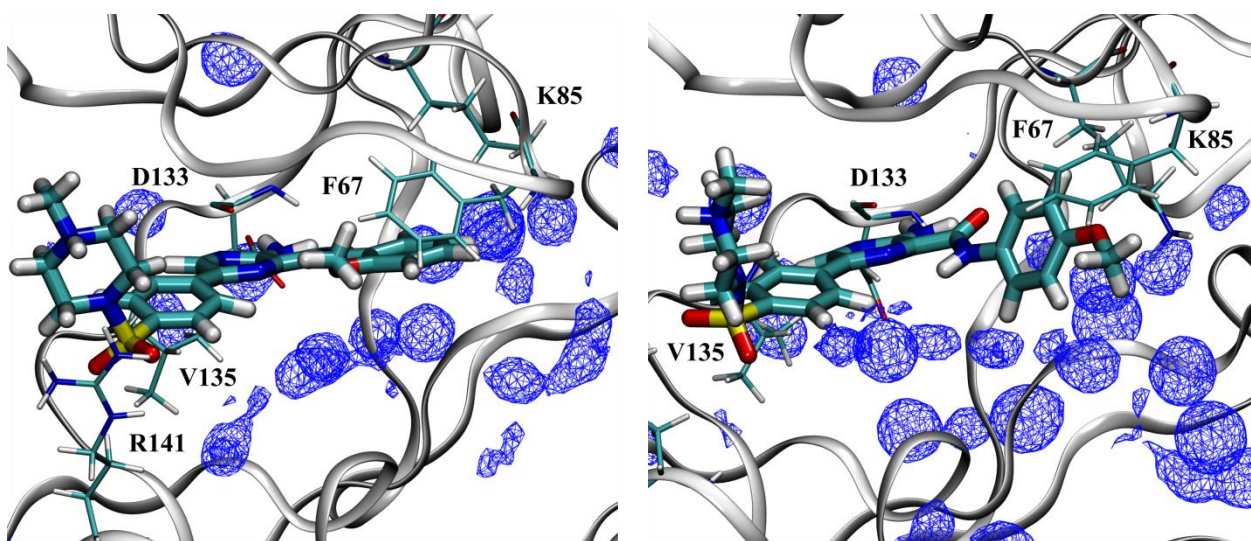
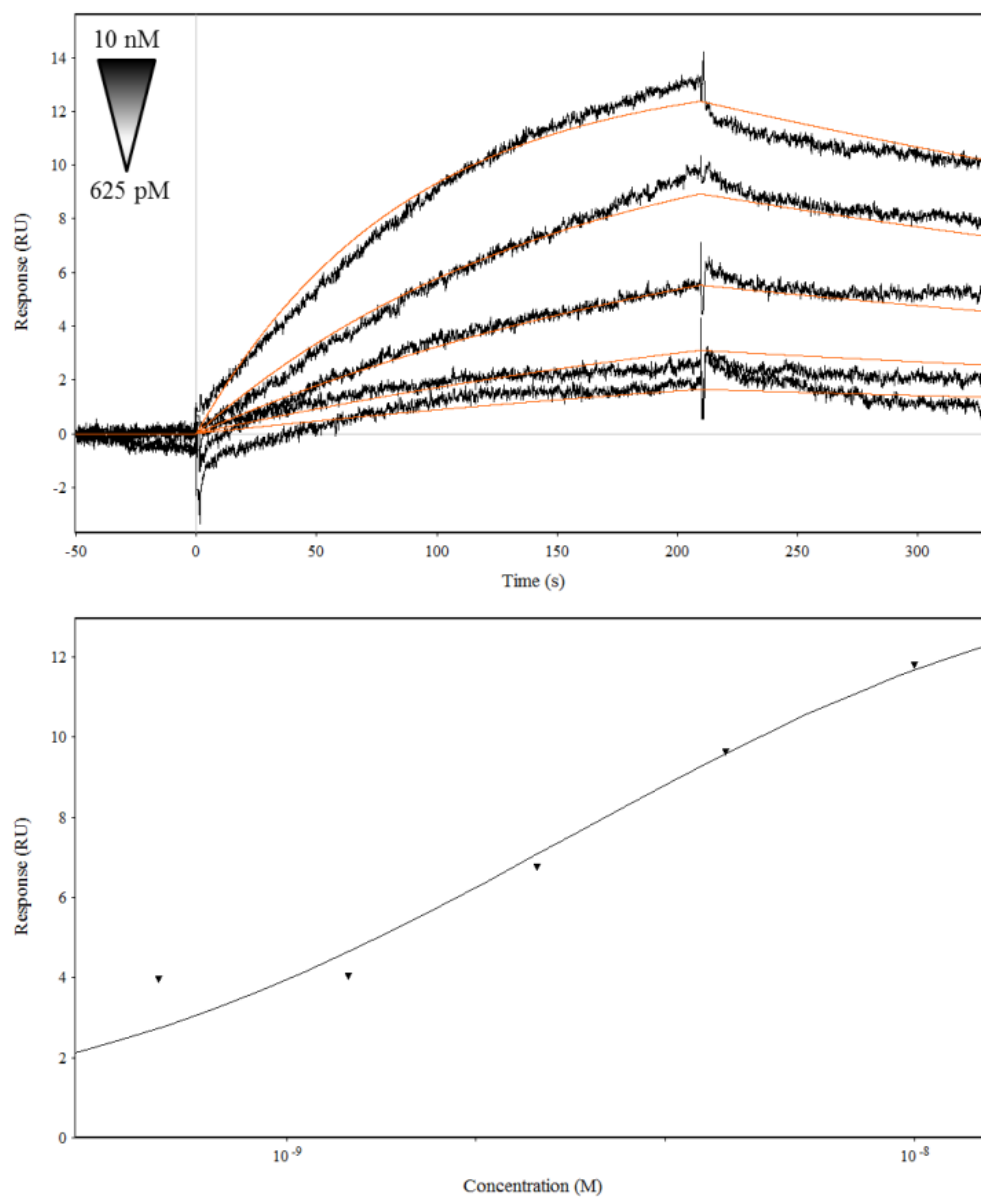


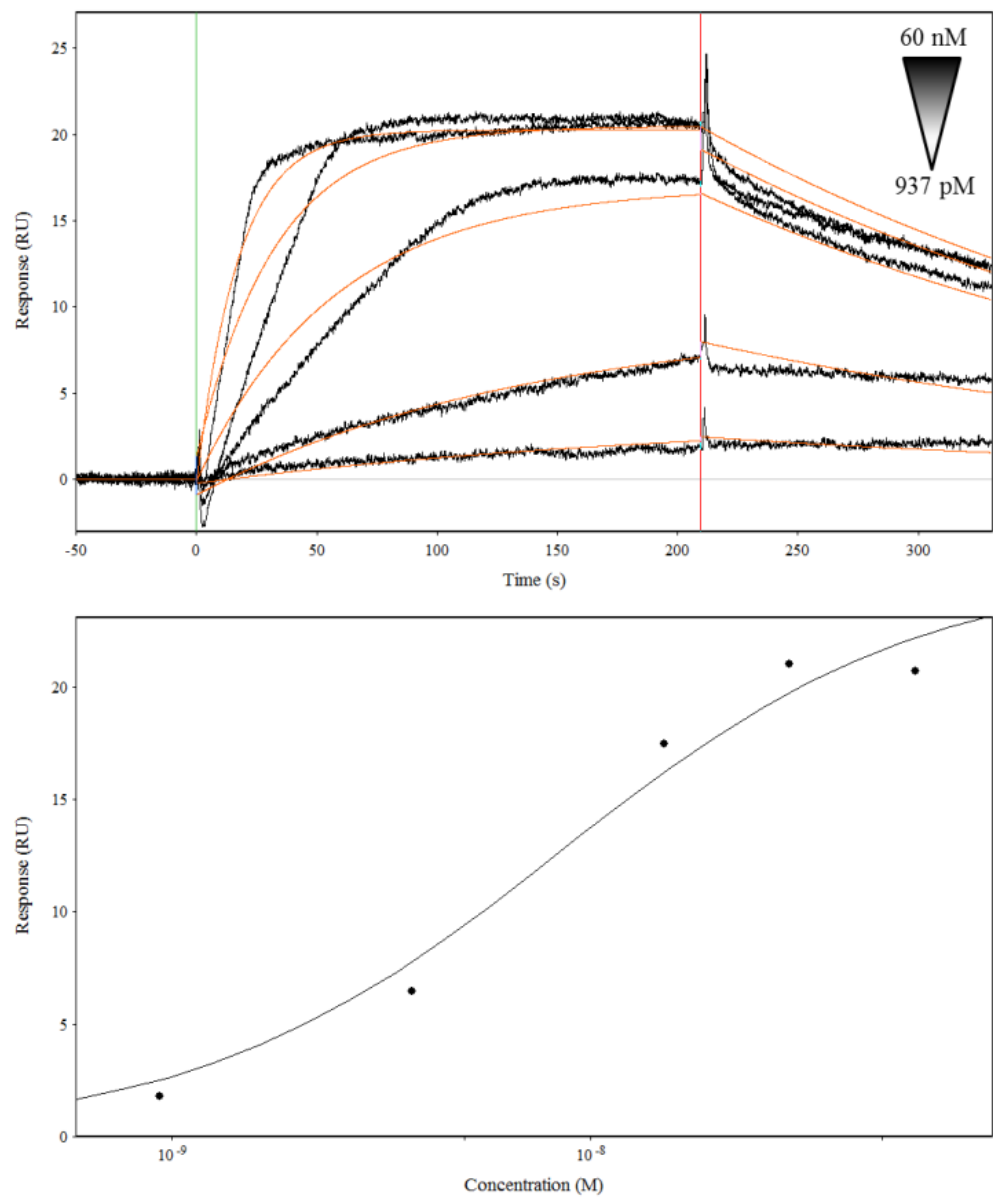
Figure A3. Probability maps representing the distribution of water molecules around compounds **5** (Left) and **7** (Right) computed when the ligands were in their bound states.

7.5. Surface plasmon resonance (SPR) affinity and binding curves

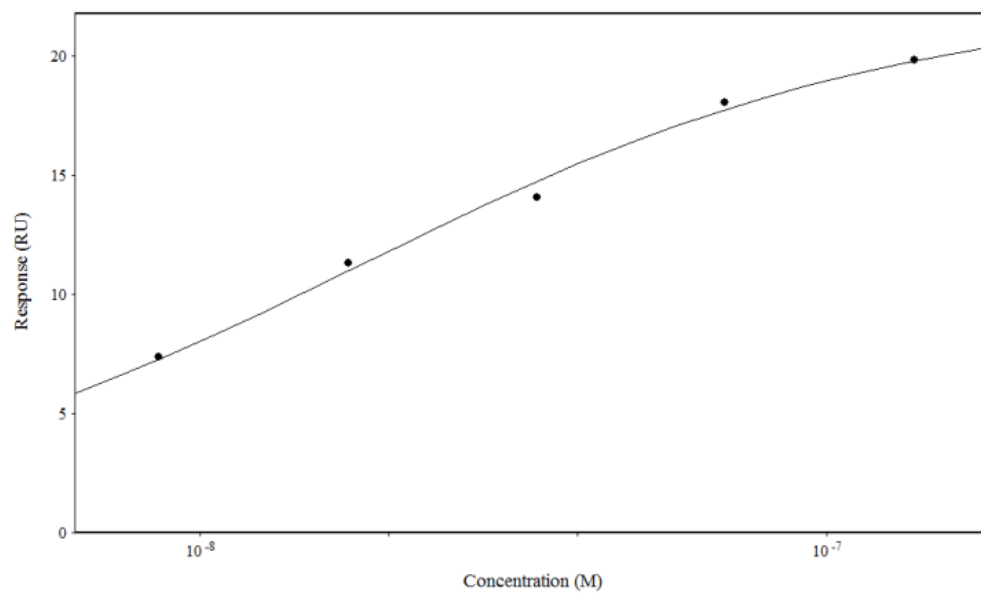
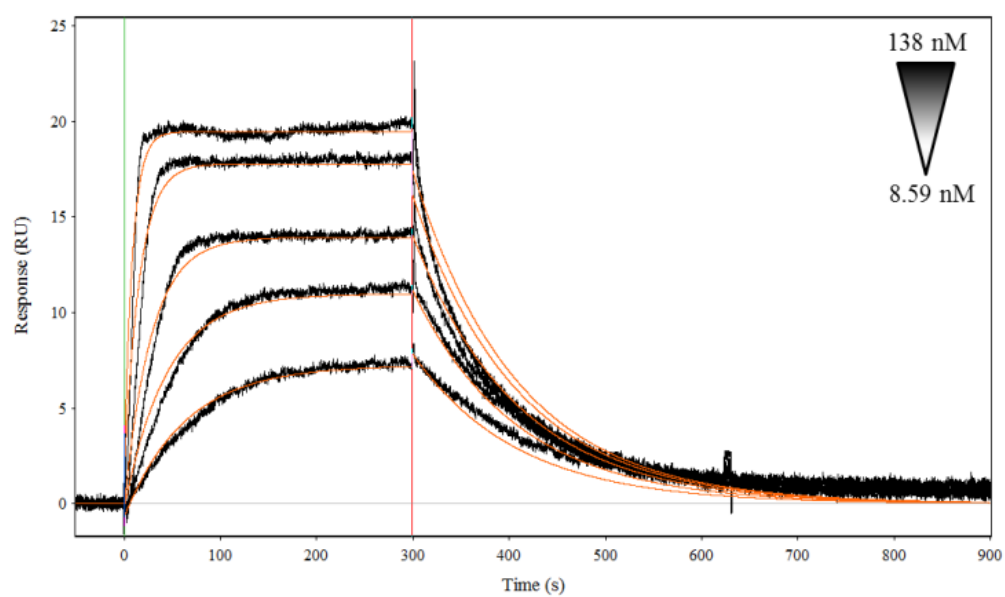
Compound 1



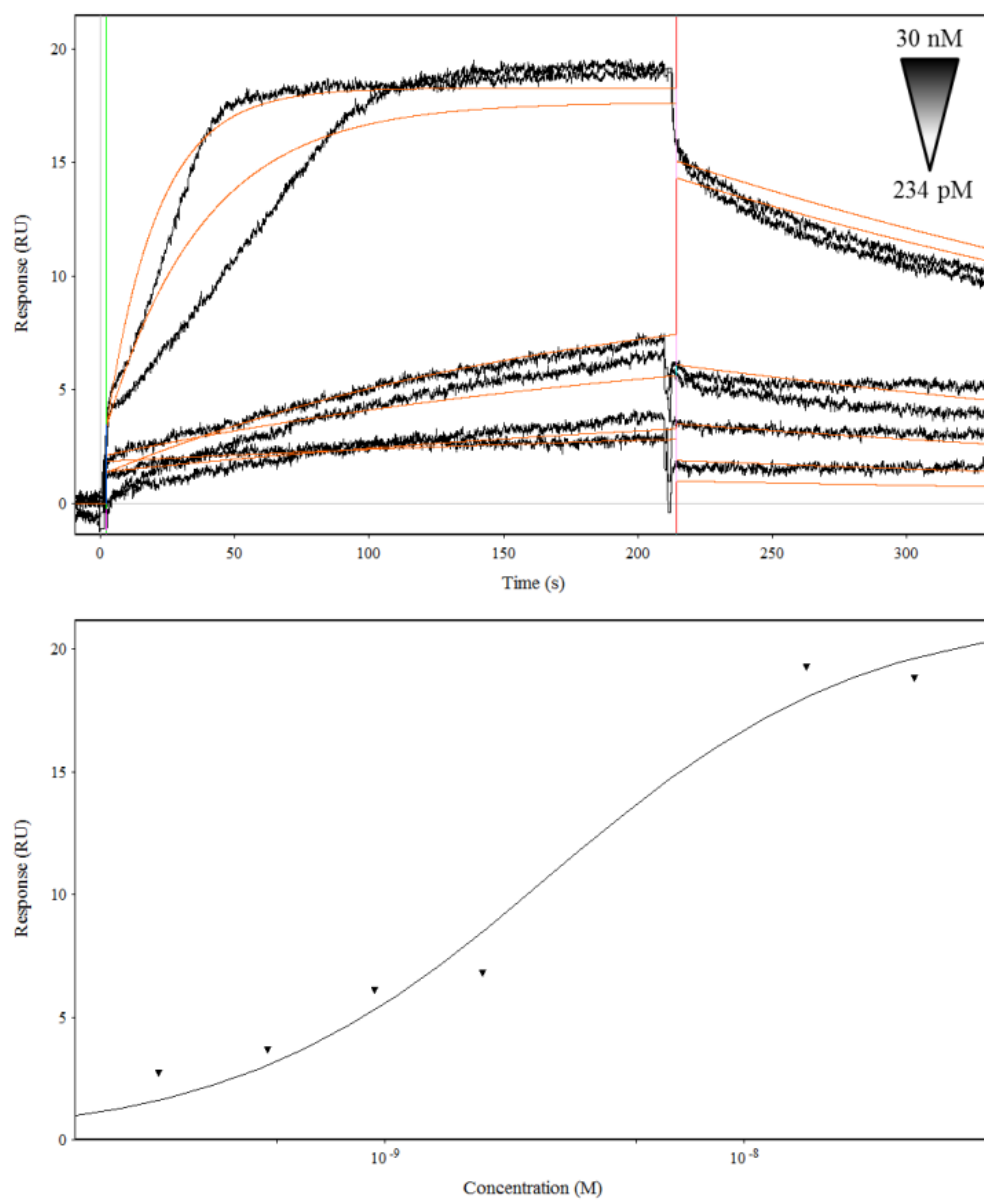
Compound 2



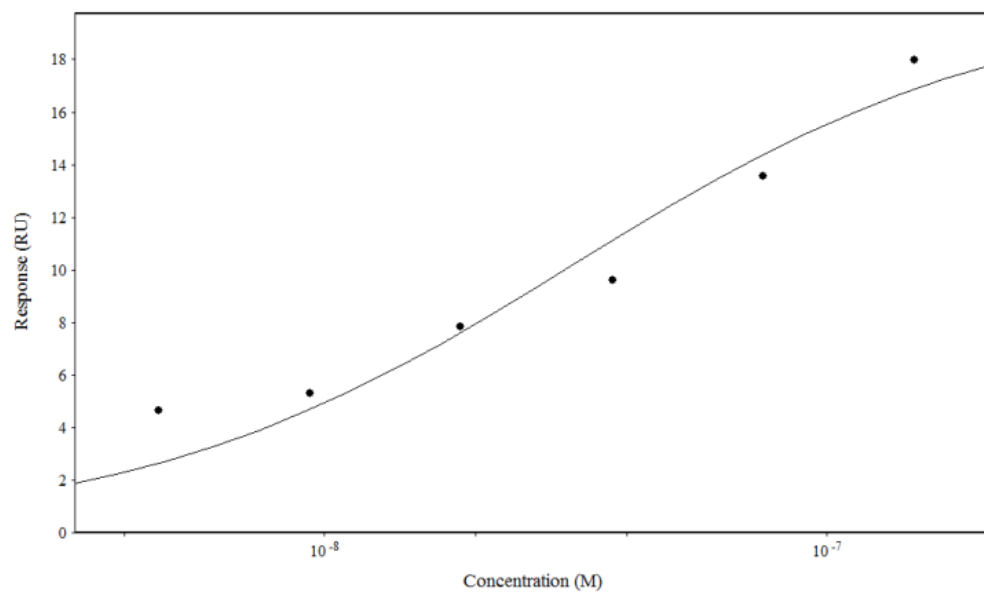
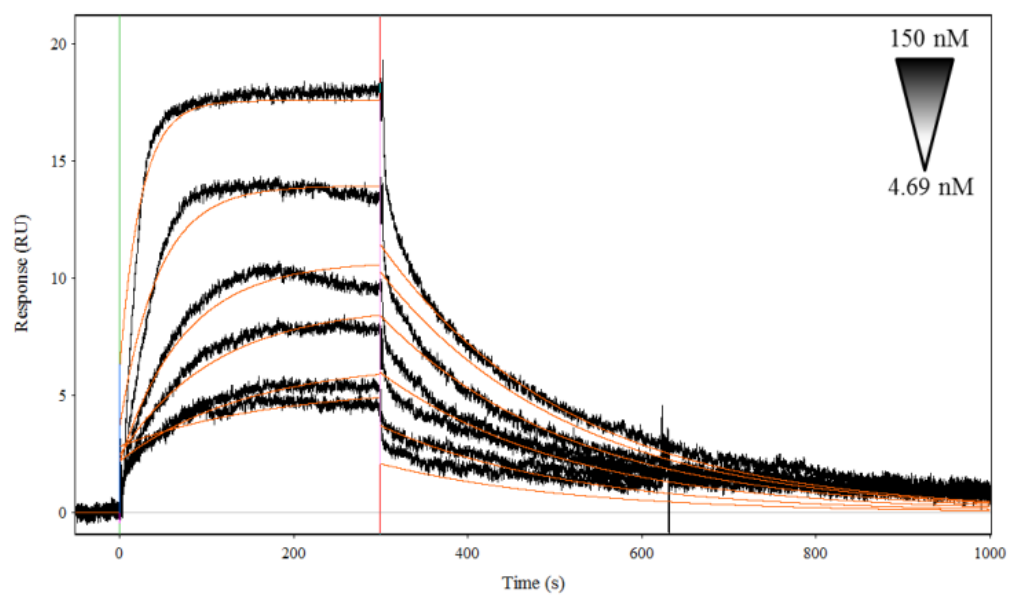
Compound 3



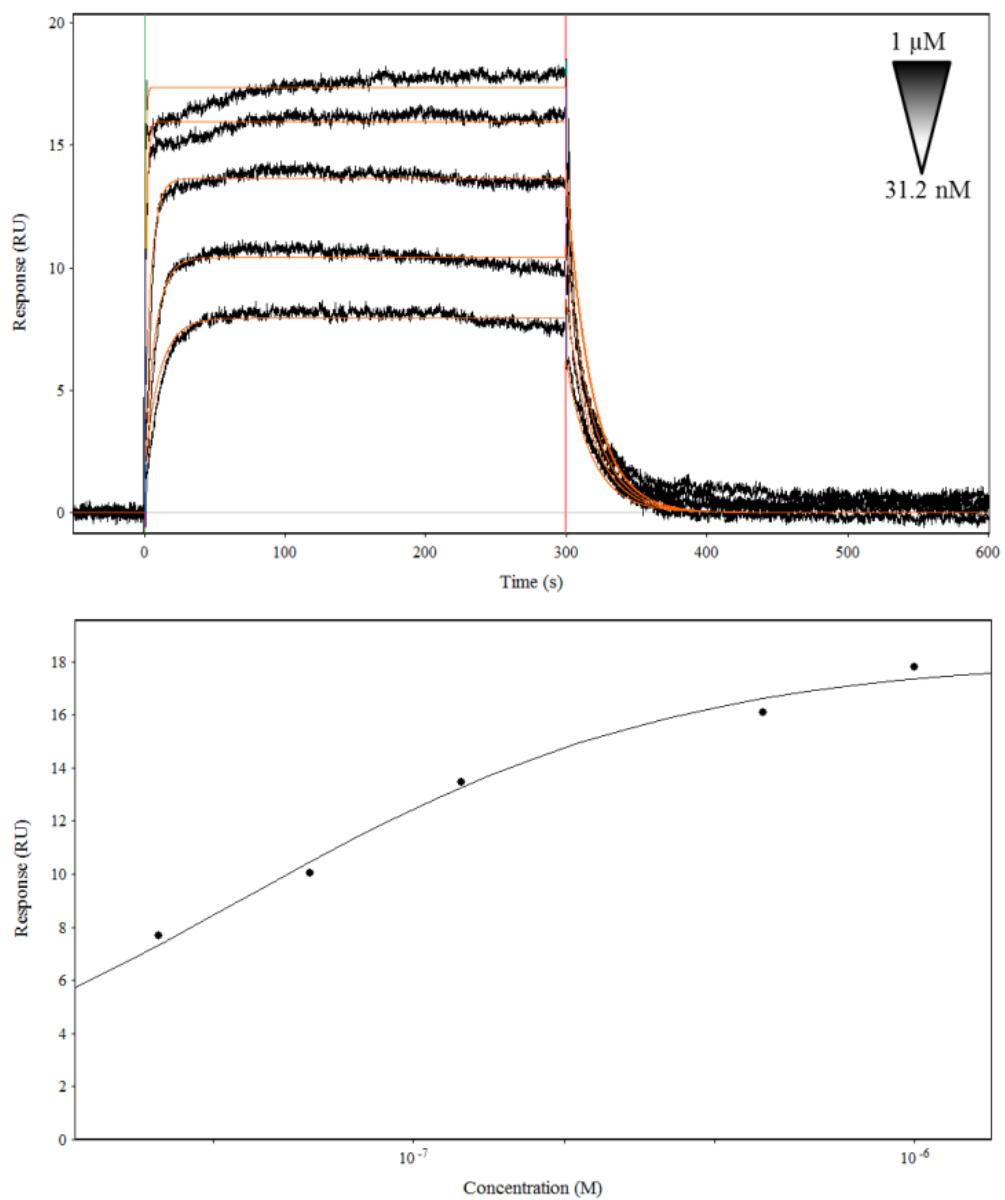
Compound 4



Compound 5



Compound 6



7.6. Further test of the force field

The performance and reliability of the force field used in our simulations is further validated in the following short subsections looking at the properties of three representative solute molecules. Benchmark data for the structure, atomic charges, and vibrational frequencies for the gas phase are provided by DFT computations carried out by the CPMD package.

7.6.1. Methane

In Figure A4, CH₄-water radial distribution function is reported. Methane-hydrogen radial distribution function shows that water hydrogen atoms move closer to methane than oxygen atoms being absent any form of repulsion between methane and other atoms in the system.

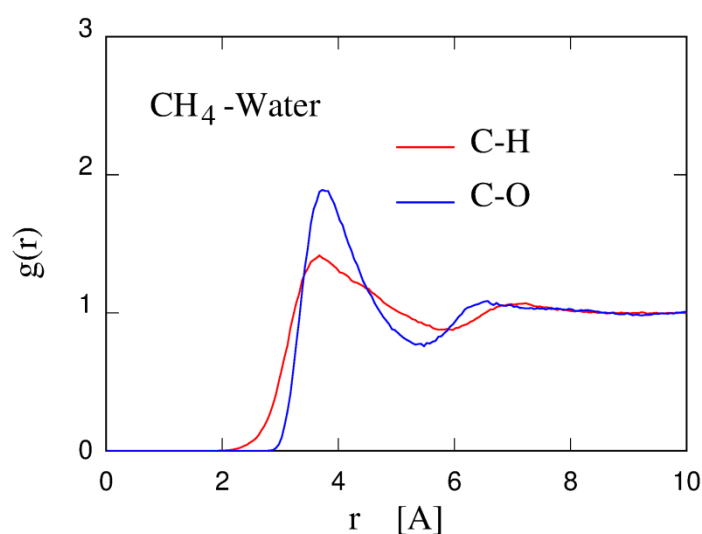


Figure A4. CH₄-water radial distribution function. CH₄-HW, CH₄-OW radial distribution functions are red and blue colored, respectively.

To comprehensively describe the dynamic behavior of methane in bulk water, the rate of diffusion was computed (see Eq. 4.25 and 4.26). The diffusion constant of methane in water resulted equal 1.71E-05 cm²/s, which is in excellent agreement with the experimental value of (1.88 ± 0.01)E-05 cm²/s.²²⁶ In the same sample, the self-diffusion of water was equal to 2.27E-05 cm²/s, whereas the experimental value is reported equal to 2.3E-05 cm²/s.

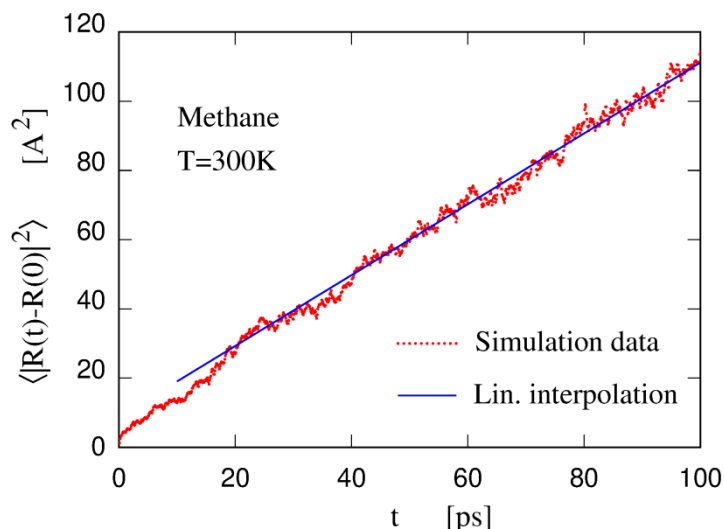


Figure A5. Mean square displacement of methane solvated in 300 SPC/Fw water molecules. The simulation data and the linear interpolation are reported in red and blue, respectively.

7.6.2. Propionic acid

We validated the force field describing the propionic acid by comparing the vibrational density of states computed by the force field and DFT extending from $\omega \sim 0$ and $\omega > 3600 \text{ cm}^{-1}$ (Fig. A6).

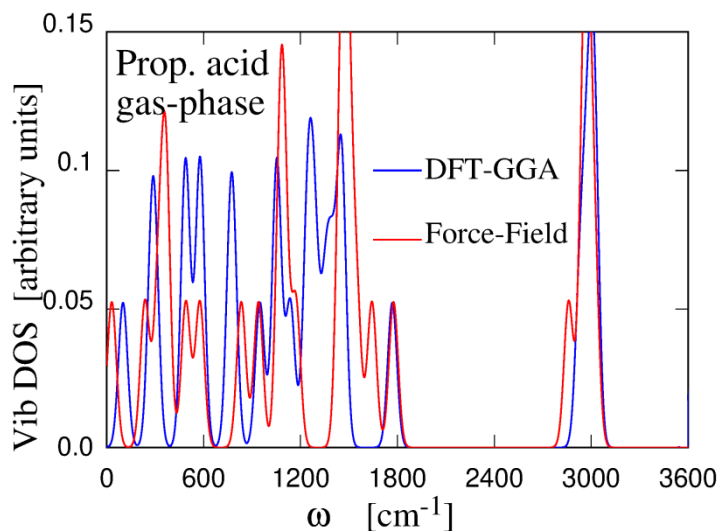


Figure A6. Comparison between the vibrational density of states computed by DFT (blue) and by the force field (red) for the neutral form of propionic acid in gas phase.

The high frequencies bands at $\omega > 3640 \text{ cm}^{-1}$ and $2928 < \omega < 3033 \text{ cm}^{-1}$ consist of O-H stretching and C-H stretching modes, respectively. C=O stretching modes are identified at $\omega = 1768 \text{ cm}^{-1}$. The band from $\omega = 1355 \text{ cm}^{-1}$ to $\omega = 1450 \text{ cm}^{-1}$ consists of H-C-H stretching modes. Various bending modes correspond to the low frequencies band from $\omega = 1050 \text{ cm}^{-1}$ to $\omega = 1250 \text{ cm}^{-1}$. At $\omega < 1000 \text{ cm}^{-1}$, mixed modes are identified.

Comparison of the ground structure modeled by the force field and DFT results in a mean square deviation (per atom), χ^2 , equal to 0.0051 \AA^2 for non-hydrogen atoms and 0.019 \AA^2 when considering all atoms.

7.6.3. Piperidine

Similarly to the propionic acid, the force field describing the piperidine was validated by comparing the vibrational density of states computed by the force field and DFT extending from $\omega \sim 0$ and $\omega > 3600 \text{ cm}^{-1}$ (Fig. A7).

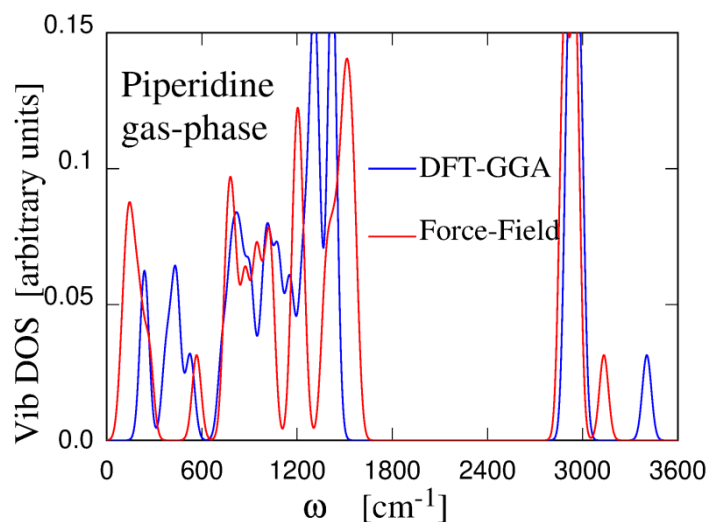


Figure A7. Comparison between the vibrational density of states computed by DFT (blue) and by the force field (red) for the neutral form of piperidine in gas phase.

The high frequencies bands at $\omega = 3600 \text{ cm}^{-1}$ and $2910 < \omega < 2984 \text{ cm}^{-1}$ consist of N-H stretching and C-H stretching modes, respectively. The band from $\omega = 1405 \text{ cm}^{-1}$ to $\omega = 1433 \text{ cm}^{-1}$ consists of H-C-H and H-N-C bending modes; the band from $\omega = 1000 \text{ cm}^{-1}$ to $\omega = 1330 \text{ cm}^{-1}$ to the H-C-C bending modes. At $\omega < 1000 \text{ cm}^{-1}$, modes are mixed.

Comparison of the ground structure modeled by the force field and DFT results in a mean square deviation (per atom), χ^2 , equal to 0.016 \AA^2 for non-hydrogen atoms and 0.035 \AA^2 when considering all atoms.

7.6.4. Nitromethane

The self-diffusion coefficient of nitromethane in water resulted to be slightly lower ($0.9045\text{E-}05 \text{ cm}^2/\text{s}$) in comparison to the equivalent result for methane (Fig. A8).

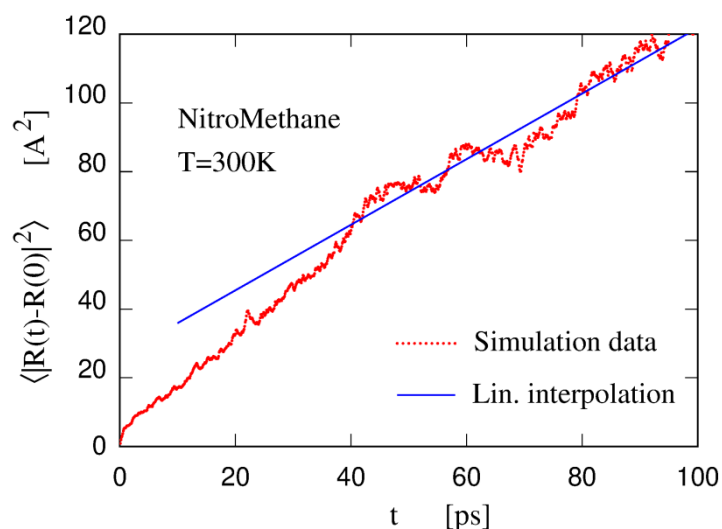


Figure A8. Mean square displacement of nitromethane solvated in 300 SPC/Fw water molecules. The simulation data and the linear interpolation are reported in red and blue, respectively.

In Figure A9, the radial distribution function of nitromethane in water is reported.

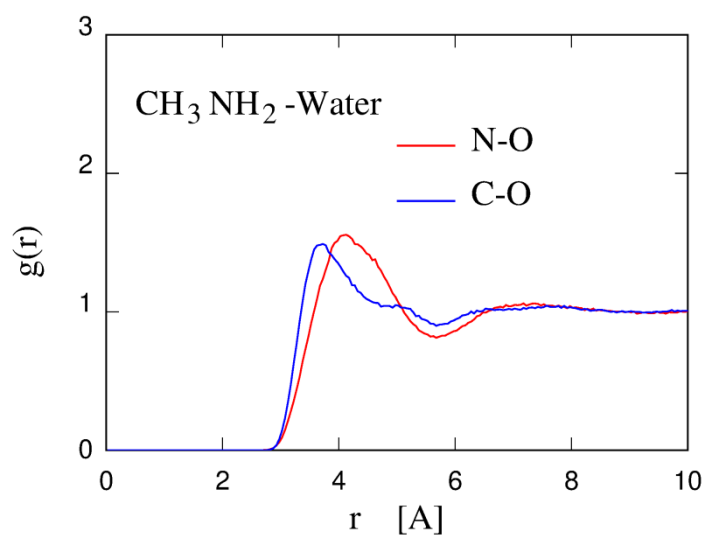


Figure A9. CH_3NH_2 -water radial distribution function. $\text{NH}_2\text{-OW}$ and $\text{CH}_3\text{-OW}$ radial distribution functions are red and blue colored, respectively.

The force field describing the nitromethane was validated by comparing the vibrational density of states computed by the force field and DFT extending from $\omega \sim 0$ and $\omega > 3111 \text{ cm}^{-1}$ (Fig. A10).

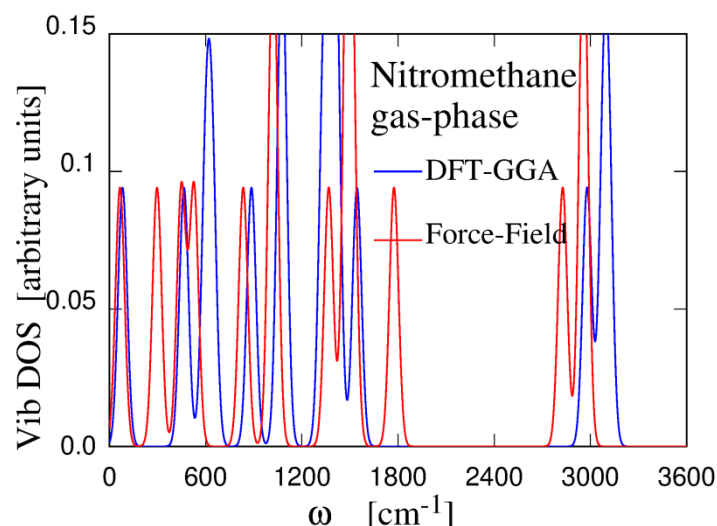


Figure A10. Comparison between the vibrational density of states computed by DFT (blue) and by the force field (red) for nitromethane in gas phase.

The high frequencies band at $3111 < \omega < 2980 \text{ cm}^{-1}$ consists of C-H stretching. At $\omega = 1540 \text{ cm}^{-1}$, the N-O stretching modes are identified. The band from $\omega = 1416 \text{ cm}^{-1}$ to $\omega = 1360 \text{ cm}^{-1}$ consists of H-C-H bending modes. At $\omega < 1360 \text{ cm}^{-1}$, modes are mixed.

Comparison of the ground structure modeled by the force field and DFT results in a mean square deviation (per atom), χ^2 , equal to 0.013 \AA^2 for non-hydrogen atoms and 0.013 \AA^2 when considering all atoms. By comparing the ground structures, the nitro group results to be slightly off-planar in the structure modeled by the force field, whereas it is perfectly planar in DFT.

7.7. Validation of the quasi-harmonic (QH) equilibrium volume

As presented in Section 4.3.5, the volume of the system was optimized by applying the quasi-harmonic (QH) approximation through small variations (1%) of the system volume. For each volume, the configuration was equilibrated and minimized by quenched MD. Each quenched configuration was then used as reference for computing the corresponding Hessian matrix. As a result, a set of volume-dependent harmonic frequencies were provided, and the minimal harmonic free energy for each temperature with respect to the volume (Fig. 4.1) was identified.

Alternatively, the equilibrium volume can be identified by performing a single NPT simulation keeping the system at the desired temperature.²²⁷

Here, we validate our approach by comparing the QH equilibrium volume in the classical approximation with the volume resulting from relatively long equilibrations at 300 K in the NPT ensemble.

Three representative systems (bulk water, nitromethane, benzene) were considered.

Systems were built with the AmberTools module named LEaP included in Amber 14.²²⁸ For the globally neutral organic small molecules, the GAFF (General Amber Force Field)⁶³ was chosen. The cubic boxes were filled with 300 SPC/Fw (flexible simple point charge)¹⁹⁷ water molecules using LEaP. Short-range electrostatic interactions were treated with the Verlet cut-off scheme, and long-range ones with the Particle Mesh Ewald (PME) method.⁷⁴ PME was chosen instead of the Ewald summation method used in the calculations presented in Chapter 4 to improve the computational performance. In both cases, the cut-off was fixed at 10 Å. Periodic boundary conditions (PBC) were applied. The time step was fixed at 1 fs. Each system was briefly minimized (50 ps) and then thermalized to 300 K in three runs using the Langevin thermostat²⁰¹ for a total of 0.3 ns of dynamics. Finally, the system volume was equilibrated at 1 atm according to the Parrinello-Rahman barostat⁷⁵ sampling the isothermal-isobaric (NPT) ensemble. All MD simulations were performed with Gromacs 4.6.1.¹⁷⁶

In Table A7, the side of the equilibrated cubic boxes obtained by the application of the QH approximation and the NPT equilibration run for the three representative systems are reported. For bulk water, convergence was achieved after 10 ns; for nitromethane and benzene after 25 ns.

As expected, the equilibrium volumes obtained by the two approaches are in good agreement.

Table A7. Comparison of cubic box sizes obtained by the QH approximation and MD runs in NPT^a

Solutes	l_{QH}	l_i	$l_{\text{f, NPT}} (5 \text{ ns})$	$l_{\text{f, NPT}} (10 \text{ ns})$	$l_{\text{f, NPT}} (25 \text{ ns})$
Water	20.9129	23.1750	20.8473	20.8217	-
Nitromethane	21.0597	21.7004	20.8544	20.8190	20.9997
Benzene	21.0816	21.3749	21.0516	20.8136	21.0119

^a Cubic box sizes, l , are expressed in Å. l_{QH} refers to the side of the cubic box at the equilibrium volume obtained by the QH approximation. l_i and $l_{\text{f, NPT}}$ refer to the side of the cubic box before and after the equilibration run in the NPT ensemble.

8. Bibliography

1. Janin, J., Protein-protein recognition. *Prog. Biophys. Mol. Biol.* **1995**, *64* (2-3), 145-166.
2. Demchenko, A. P., Recognition between flexible protein molecules: induced and assisted folding. *J. Mol. Recognit.* **2001**, *14* (1), 42-61.
3. Du, X.; Li, Y.; Xia, Y.-L.; Ai, S.-M.; Liang, J.; Sang, P.; Ji, X.-L.; Liu, S.-Q., Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Mol. Sci.* **2016**, *17* (2), 144.
4. Gabanyi, M. J.; Berman, H. M., Protein Structure Annotation Resources. *Methods Mol. Biol.* **2015**, *1261*, 3-20.
5. Fisher, E., The influence of configuration on enzyme activity.(Translated from German). *Dtsch. Chem. Ges.* **1894**, *27*, 2984-2993.
6. Koshland, D., Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA* **1958**, *44* (2), 98-104.
7. Tobi, D.; Bahar, I., Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. USA* **2005**, *102* (52), 18908-18913.
8. Acuner Ozbabacan, S. E.; Gursoy, A.; Keskin, O.; Nussinov, R., Conformational ensembles, signal transduction and residue hot spots: application to drug discovery. *Curr. Opin. Drug Discov. Devel.* **2010**, *13* (5), 527-537.
9. Csermely, P.; Palotai, R.; Nussinov, R., Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **2010**, *35* (10), 539-546.
10. Changeux, J.-P.; Edelstein, S., Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol. Rep.* **2011**, *3*, 19.
11. Boehr, D. D.; Nussinov, R.; Wright, P. E., The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5* (11), 789-796.
12. Copeland, R. A., Conformational adaptation in drug-target interactions and residence time. *Future Med. Chem.* **2011**, *3* (12), 1491-1501.
13. Gilson, M. K.; Zhou, H.-X., Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*.
14. Gibbs, J. W., A Method of Geometrical Representation of the Thermodynamic Properties by Means of Surfaces. *Transactions of Connecticut Academy of Arts and Sciences* **1873**, 382-404.
15. Martin, S. F.; Clements, J. H., Correlating structure and energetics in protein-ligand interactions: paradigms and paradoxes. *Annu. Rev. Biochem.* **2013**, *82*, 267-293.
16. Keseru, G.; Swinney, D. C., Thermodynamics and Binding Kinetics in Drug Discovery. *Thermodynamics and Kinetics of Drug Binding* **2015**, 313-329.
17. Bongrand, P., Ligand-receptor interactions. *Rep. Prog. Phys.* **1999**, *62* (6), 921.

18. Xie, Y.-H.; Tao, Y.; Liu, S.-Q., 153 Wonderful roles of the entropy in protein dynamics, binding and folding. *J. Biomol. Struct. Dyn.* **2013**, *31* (sup1), 98-100.
19. Held, M.; Metzner, P.; Prinz, J.-H.; Noé, F., Mechanisms of protein-ligand association and its modulation by protein mutations. *Biophys. J.* **2011**, *100* (3), 701-710.
20. Arrhenius, S., Concerning the heat of dissociation and the influence of the temperature on the degree of dissociation of electrolytes. *Z. Phys. Chem.* **1889**, *4*, 96-116.
21. Laidler, K. J.; King, M. C., Development of transition-state theory. *The Journal of Physical Chemistry* **1983**, *87* (15), 2657-2664.
22. Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J., Current status of transition-state theory. *The Journal of Physical Chemistry* **1996**, *100* (31), 12771-12800.
23. Schuetz, D. A.; de Witte, W. E. A.; Wong, Y. C.; Knasmueller, B.; Richter, L.; Kokh, D. B.; Sadiq, S. K.; Bosma, R.; Nederpelt, I.; Heitman, L. H., Kinetics for drug discovery: an industry-driven effort to target drug residence time. *Drug Discov. Today* **2017**, *22* (6), 896-911.
24. Copeland, R. A.; Pompliano, D. L.; Meek, T. D., Drug–target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* **2006**, *5*, 730-739.
25. Bosch, F.; Rosich, L., The Contributions of Paul Ehrlich to Pharmacology: A Tribute on the Occasion of the Centenary of His Nobel Prize. *Pharmacology* **2008**, *82* (3), 171-179.
26. Langley, J. N., On the reaction of cells and of nerve-endings to certain poisons, chiefly as regards the reaction of striated muscle to nicotine and to curari. *J. Physiol.* **1905**, *33* (4-5), 374-413.
27. De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A., Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035-4061.
28. Copeland, R. A., The drug-target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov.* **2016**, *15* (2), 87-95.
29. Tonge, P. J., Drug–target kinetics in drug discovery. *ACS Chem. Neurosci.* **2017**, *9* (1), 29-39.
30. Dahl, G.; Akerud, T., Pharmacokinetics and the drug–target residence time concept. *Drug Discov. Today* **2013**, *18* (15-16), 697-707.
31. Folmer, R. H., Drug target residence time: a misleading concept. *Drug Discov. Today* **2018**, *23* (1), 12-16.
32. Freire, E.; Mayorga, O. L.; Straume, M., Isothermal titration calorimetry. *Anal. Chem.* **1990**, *62* (18), 950A-959A.
33. Serdyuk, I. N.; Zaccai, J.; Zaccai, N. R., Isothermal titration calorimetry. In *Methods in Molecular Biophysics: Structure, Dynamics, Function*, Cambridge University Press: Cambridge, 2007; pp 221-233.
34. de Mol, N. J.; Fischer, M. J., Kinetic and thermodynamic analysis of ligand-receptor interactions: SPR applications in drug development. *Handbook of Surface Plasmon Resonance* **2008**, 123-172.
35. Atkins, P.; De Paula, J.; Keeler, J., *Atkins' physical chemistry*. Oxford university press: 2018.

36. Ingólfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J., The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, 4 (3), 225-248.
37. Chandler, D., Introduction to Modern Statistical Mechanics. *Oxford University Press* **1987**
38. Allen, M. P.; Tildesley, D. J., *Computer Simulation of Liquids: Second Edition*. OUP Oxford: 2017.
39. Meyer, K.; Hall, G.; Offin, D., *Introduction to Hamiltonian Dynamical Systems and the N-body Problem*. Springer Science & Business Media: 2008; Vol. 90.
40. Frenkel, D.; Smit, B., *Understanding Molecular Simulation*. Academic Press: 2001; p 638.
41. Born, M.; Oppenheimer, R., Zur Quantentheorie der Molekeln. *Ann. Phys.* **1927**, 389 (20), 457-484.
42. Born, M.; Huang, K., *Dynamical theory of crystal lattices*. Clarendon press: 1954.
43. Pauling, L.; Wilson, E. B., *Introduction to quantum mechanics with applications to chemistry*. Courier Corporation: 2012.
44. Marcelin, M. R., Contribution à l'étude de la cinétique physico-chimique. *Eur. Phys. J. H* **1915**, 9 (3), 120-231.
45. Lewars, E. G., Computational Chemistry. *Springer Netherlands* **2011**, 664.
46. Morse, P. M., Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels. *Phys. Rev.* **1929**, 34 (1), 57-64.
47. Beyer, M. K., The mechanical strength of a covalent bond calculated by density functional theory. *J. Chem. Phys.* **2000**, 112 (17), 7307-7312.
48. Jones, J. E.; Sc., D., On the determination of molecular fields. —II. From the equation of state of a gas. *Proc. R. Soc. Lond. A* **1924**, 106 (738), 463-477.
49. Lorentz, H. A., Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. *Ann. Phys.* **1881**, 248 (1), 127-136.
50. van Gunsteren, W. F.; Weiner, P. K.; Wilkinson, A. J., Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications *Springer Netherlands* **1997**, 3, 618.
51. Salomon-Ferrer, R.; Case, D. A.; Walker, R. C., An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, 3 (2), 198-210.
52. Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, 30 (10), 1545-1614.
53. Jorgensen, W. L.; Tirado-Rives, J., The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, 110 (6), 1657-1666.

54. Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R., GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* **1995**, *91* (1), 43-56.
55. Lopes, P. E. M.; Guvench, O.; MacKerell, A. D., Current Status of Protein Force Fields for Molecular Dynamics. *Methods Mol. Biol.* **2015**, *1215*, 47-71.
56. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P., A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106* (3), 765-784.
57. Hagler, A. T.; Huler, E.; Lifson, S., Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.* **1974**, *96* (17), 5319-5327.
58. Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A., An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7* (2), 230-252.
59. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple AMBER force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65* (3), 712-725.
60. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179-5197.
61. Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y., Strike a balance: Optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *J. Comput. Chem.* **2006**, *27* (6), 781-790.
62. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696-3713.
63. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25* (9), 1157-1174.
64. Marx, D.; Hutter, J., *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge University Press: Cambridge, 2009.
65. Car, R.; Parrinello, M., Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.* **1985**, *55* (22), 2471-2474.
66. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865-3868.
67. Kim, K.; Jordan, K. D., Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *The Journal of Physical Chemistry* **1994**, *98* (40), 10089-10094.
68. Ashcroft, N. W.; Mermin, D., *Solid State Physics*. Holt, Rinehart and Winston **1976**.
69. Srivastava, G. P., *The Physics of Phonons*. Taylor & Francis: 1990.

70. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327-341.
71. Verlet, L., Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98-103.
72. Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R., A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **1982**, *76* (1), 637-649.
73. Leeuw, S. W. d.; Perram, J. W.; Smith, E. R., Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* **1980**, *373* (1752), 27-56.
74. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089-10092.
75. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52* (12), 7182-7190.
76. Nosé, S., A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81* (1), 511-519.
77. Hoover, W. G., Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31* (3), 1695-1697.
78. Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81* (8), 3684-3690.
79. Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126* (1), 014101.
80. Hünenberger, P., Thermostat Algorithms for Molecular Dynamics Simulations in Advanced Computer Simulation. *Springer-Verlag Berlin Heidelberg* **2005**, *173*, 105-149.
81. Hammersley, J., Monte Carlo Methods. *Springer Netherlands* **1964**.
82. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21* (6), 1087-1092.
83. Arrhenius, S., Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissoziationsgrad der Elektrolyte. In *Z. Phys. Chem.*, 1889; Vol. 4U, pp 96-116.
84. Truhlar, D. G.; Garrett, B. C., Variational Transition State Theory. *Annu. Rev. Phys. Chem.* **1984**, *35* (1), 159-189.
85. Zwanzig, R. W., High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420-1426.
86. Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22* (2), 245-268.
87. Kirkwood, J. G., Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3* (5), 300-313.

88. Christ, C. D.; Mark, A. E.; van Gunsteren, W. F., Basic ingredients of free energy calculations: A review. *J. Comput. Chem.* **2010**, *31* (8), 1569-1582.
89. Squire, D. R.; Hoover, W. G., Monte Carlo Simulation of Vacancies in Rare-Gas Crystals. *J. Chem. Phys.* **1969**, *50* (2), 701-706.
90. Torrie, G. M.; Valleau, J. P., Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23* (2), 187-199.
91. Valleau, J. P.; Card, D. N., Monte Carlo Estimation of the Free Energy by Multistage Sampling. *J. Chem. Phys.* **1972**, *57* (12), 5457-5462.
92. Rosso, L.; Mináry, P.; Zhu, Z.; Tuckerman, M. E., On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles. *J. Chem. Phys.* **2002**, *116* (11), 4389-4402.
93. Chipot, C.; Pohorille, A., *Free energy calculations*. Springer: 2007.
94. Simonson, T.; Archontis, G.; Karplus, M., Free energy simulations come of age: protein– ligand recognition. *Acc. Chem. Res.* **2002**, *35* (6), 430-437.
95. Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A., The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **1997**, *72* (3), 1047-1069.
96. De Jong, D. H.; Schafer, L. V.; De Vries, A. H.; Marrink, S. J.; Berendsen, H. J.; Grubmüller, H., Determining equilibrium constants for dimerization reactions from molecular dynamics simulations. *J. Comput. Chem.* **2011**, *32* (9), 1919-1928.
97. Abrams, C.; Bussi, G., Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **2014**, *16* (1), 163-199.
98. Montalvo-Acosta, J. J.; Cecchini, M., Computational Approaches to the Chemical Equilibrium Constant in Protein-ligand Binding. *Mol. Inform.* **2016**, *35* (11-12), 555-567.
99. Mobley, D. L.; Gilson, M. K., Predicting binding free energies: Frontiers and benchmarks. *Annu. Rev. Biophys.* **2017**, *46*, 531-558.
100. Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J., Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *J. Chem. Phys.* **1988**, *89* (6), 3742-3746.
101. Hermans, J.; Wang, L., Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding. Application to a complex of benzene and mutant T4 lysozyme. *J. Am. Chem. Soc.* **1997**, *119* (11), 2707-2714.
102. Hamelberg, D.; McCammon, J. A., Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. *J. Am. Chem. Soc.* **2004**, *126* (24), 7683-7689.
103. Wang, J.; Deng, Y.; Roux, B., Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys. J.* **2006**, *91* (8), 2798-2814.
104. Helms, V.; Wade, R. C., Computational alchemy to calculate absolute protein– ligand binding free energy. *J. Am. Chem. Soc.* **1998**, *120* (12), 2710-2713.

105. Woods, C. J.; Malaisree, M.; Hannongbua, S.; Mulholland, A. J., A water-swap reaction coordinate for the calculation of absolute protein–ligand binding free energies. *J. Chem. Phys.* **2011**, *134* (5), 02B611.
106. Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C., Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.* **2016**, *7* (1), 207-218.
107. Rocchia, W.; Masetti, M.; Cavalli, A., Enhanced sampling methods in drug design. *Physico-Chemical and Computational Approaches to Drug Discovery. The Royal Society of Chemistry* **2012**, 273-301.
108. Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A., The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13* (8), 1011-1021.
109. Isralewitz, B.; Gao, M.; Schulten, K., Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* **2001**, *11* (2), 224-230.
110. Grubmüller, H.; Heymann, B.; Tavan, P., Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* **1996**, *271* (5251), 997-999.
111. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **2002**, *99* (20), 12562-12566.
112. Pietrucci, F.; Marinelli, F.; Carloni, P.; Laio, A., Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations. *J. Am. Chem. Soc.* **2009**, *131* (33), 11811-11818.
113. Limongelli, V.; Bonomi, M.; Marinelli, L.; Gervasio, F. L.; Cavalli, A.; Novellino, E.; Parrinello, M., Molecular basis of cyclooxygenase enzymes (COXs) selective inhibition. *Proc. Natl. Acad. Sci. USA* **2010**, *107* (12), 5411-5416.
114. Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M., Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. USA* **2015**, *112* (5), E386-E391.
115. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314* (1-2), 141-151.
116. Liu, P.; Kim, B.; Friesner, R. A.; Berne, B., Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. USA* **2005**, *102* (39), 13749-13754.
117. Mollica, L.; Decherchi, S.; Zia, S. R.; Gaspari, R.; Cavalli, A.; Rocchia, W., Kinetics of protein–ligand unbinding via smoothed potential molecular dynamics simulations. *Sci. Rep.* **2015**, *5*, 11539.
118. Mollica, L.; Theret, I.; Antoine, M.; Perron-Sierra, F.; Charton, Y.; Fourquez, J. M.; Wierzbicki, M.; Boutin, J. A.; Ferry, G.; Decherchi, S.; Bottegoni, G.; Ducrot, P.; Cavalli, A., Molecular Dynamics Simulations and Kinetic Measurements to Estimate and Predict Protein-Ligand Residence Times. *J. Med. Chem.* **2016**, *59* (15), 7167-7176.
119. Bernetti, M.; Cavalli, A.; Mollica, L., Protein-ligand (un)binding kinetics as a new paradigm for drug discovery at the crossroad between experiments and modelling. *MedChemComm* **2017**, *8* (3), 534-550.
120. Bruce, N. J.; Ganotra, G. K.; Kokh, D. B.; Sadiq, S. K.; Wade, R. C., New approaches for computing ligand–receptor binding kinetics. *Curr. Opin. Struct. Biol.* **2018**, *49*, 1-10.

121. Hoover, W. G., Gray, S. G., Johnson, K. W. , Thermodynamic Properties of the Fluid and Solid Phases for Inverse Power Potentials. *J. Chem. Phys.* **1971**, *55* (3), 1128-1136.
122. Frenkel, D., Ladd A. J. C., New Monte Carlo method to compute the free energy of arbitrary solids. Application to the fcc and hcp phases of hard spheres. *J. Chem. Phys.* **1984**, *81* (7), 3188-3193.
123. Johnson, J. K.; Zollweg, J. A.; Gubbins, K. E., The Lennard-Jones equation of state revisited. *Mol. Phys.* **1993**, *78* (3), 591-618.
124. Stoessel, J. P.; Nowak, P., Absolute free energies in biomolecular systems. *Macromolecules* **1990**, *23* (7), 1961-1965.
125. Gō, N.; Scheraga, H. A., Analysis of the contribution of internal vibrations to the statistical weights of equilibrium conformations of macromolecules. *J. Chem. Phys.* **1969**, *51* (11), 4751-4767.
126. Karplus, M.; Kushick, J. N., Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981**, *14* (2), 325-332.
127. Ma, B.; Tsai, C. J.; Nussinov, R., A systematic study of the vibrational free energies of polypeptides in folded and random states. *Biophys. J.* **2000**, *79* (5), 2739-2753.
128. Hagler, A.; Stern, P.; Sharon, R.; Becker, J.; Naider, F., Computer simulation of the conformational properties of oligopeptides. Comparison of theoretical methods and analysis of experimental results. *J. Am. Chem. Soc.* **1979**, *101* (23), 6842-6852.
129. Meirovitch, H., Calculation of the Free Energy and the Entropy of Macromolecular Systems by Computer Simulation. In *Reviews in Computational Chemistry*, Lipkowitz, K. B.; Boyd, D. B., Eds. 2007.
130. Cheluvaraja, S., Meirovitch, H., Calculation of the entropy and free energy of peptides by molecular dynamics simulations using the hypothetical scanning molecular dynamics method. *J. Chem. Phys.* **2006**, *125* (2), 024905.
131. White, R. P., Meirovitch, H., Lower and upper bounds for the absolute free energy by the hypothetical scanning Monte Carlo method: Application to liquid argon and water. *J. Chem. Phys.* **2004**, *121* (22), 10889-10904.
132. Szarecka, A.; White, R. P.; Meirovitch, H., Absolute entropy and free energy of fluids using the hypothetical scanning method. I. Calculation of transition probabilities from local grand canonical partition functions. *J. Chem. Phys.* **2003**, *119* (23), 12084-12095.
133. Head, M. S.; Given, J. A.; Gilson, M. K., "Mining Minima": Direct Computation of Conformational Free Energy. *The Journal of Physical Chemistry A* **1997**, *101* (8), 1609-1618.
134. Ytreberg, F. M.; Zuckerman, D. M., Simple estimation of absolute free energies for biomolecules. *J Chem Phys* **2006**, *124* (10), 104105.
135. Tyka, M. D.; Clarke, A. R.; Sessions, R. B., An efficient, path-independent method for free-energy calculations. *The Journal of Physical Chemistry B* **2006**, *110* (34), 17212-20.
136. Tyka, M. D.; Sessions, R. B.; Clarke, A. R., Absolute Free-Energy Calculations of Liquids Using a Harmonic Reference State. *The Journal of Physical Chemistry B* **2007**, *111* (32), 9571-9580.

137. Strajbl, M.; Sham, Y. Y.; Villà, J.; Chu, Z. T.; Warshel, A., Calculations of Activation Entropies of Chemical Reactions in Solution. *The Journal of Physical Chemistry B* **2000**, *104* (18), 4578-4584.
138. Villà, J.; Štrajbl, M.; Glennon, T. M.; Sham, Y. Y.; Chu, Z. T.; Warshel, A., How important are entropic contributions to enzyme catalysis? *Proceedings of the National Academy of Sciences of the United States of America* **2000**, *97* (22), 11899-11904.
139. Cecchini, M.; Krivov, S. V.; Spichty, M.; Karplus, M., Calculation of free-energy differences by confinement simulations. Application to peptide conformers. *The Journal of Physical Chemistry B* **2009**, *113* (29), 9728-9740.
140. Esque, J.; Cecchini, M., Accurate Calculation of Conformational Free Energy Differences in Explicit Water: The Confinement–Solvation Free Energy Approach. *The Journal of Physical Chemistry B* **2015**, *119* (16), 5194-5207.
141. Henchman, R. H., Partition function for a simple liquid using cell theory parametrized by computer simulation. *J. Chem. Phys.* **2003**, *119* (1), 400-406.
142. Barker, J. A., *Lattice theories of the liquid state*. Pergamon Press: 1963; Vol. 1.
143. Henchman, R. H., Free energy of liquid water from a computer simulation via cell theory. *J. Chem. Phys.* **2007**, *126* (6), 064504.
144. Klefas-Stennett, M.; Henchman, R. H., Classical and quantum Gibbs free energies and phase behavior of water using simulation and cell theory. *The Journal of Physical Chemistry B* **2008**, *112* (32), 9769-9776.
145. Zielkiewicz, J., Entropy of water calculated from harmonic approximation: Estimation of the accuracy of method. *J. Chem. Phys.* **2008**, *128* (19), 196101.
146. Habershon, S.; Manolopoulos, D. E., Free energy calculations for a flexible water model. *Phys. Chem. Chem. Phys.* **2011**, *13* (44), 19714-27.
147. Habershon, S.; Markland, T. E.; Manolopoulos, D. E., Competing quantum effects in the dynamics of a flexible water model. *J. Chem. Phys.* **2009**, *131* (2), 024501.
148. Li, L.; Totton, T.; Frenkel, D., Computational methodology for solubility prediction: Application to the sparingly soluble solutes. *J. Chem. Phys.* **2017**, *146* (21), 214110.
149. Li, L.; Totton, T.; Frenkel, D., Computational methodology for solubility prediction: Application to sparingly soluble organic/inorganic materials. *J. Chem. Phys.* **2018**, *149* (5), 054102.
150. Walkup, G. K.; You, Z.; Ross, P. L.; Allen, E. K. H.; Daryaei, F.; Hale, M. R.; O'Donnell, J.; Ehmann, D. E.; Schuck, V. J. A.; Buurman, E. T.; Choy, A. L.; Hajec, L.; Murphy-Benenato, K.; Marone, V.; Patey, S. A.; Grosser, L. A.; Johnstone, M.; Walker, S. G.; Tonge, P. J.; Fisher, S. L., Translating slow-binding inhibition kinetics into cellular and in vivo effects. *Nat. Chem. Biol.* **2015**, *11* (6), 416-423.
151. Muller, P. Y.; Milton, M. N., The determination and interpretation of the therapeutic index in drug development. *Nat. Rev. Drug Discov.* **2012**, *11*, 751-761.
152. Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E., Molecular determinants of drug-receptor binding kinetics. *Drug Discov. Today* **2013**, *18* (13-14), 667-673.

153. Shukla, D.; Hernández, C. X.; Weber, J. K.; Pande, V. S., Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Acc. Chem. Res.* **2015**, *48* (2), 414-422.
154. Buch, I.; Giorgino, T.; De Fabritiis, G., Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **2011**, *108* (25), 10184-10189.
155. Tiwary, P.; Parrinello, M., From metadynamics to dynamics. *Phys. Rev. Lett.* **2013**, *111* (23), 230602.
156. Casasnovas, R.; Limongelli, V.; Tiwary, P.; Carloni, P.; Parrinello, M., Unbinding Kinetics of a p38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. *J. Am. Chem. Soc.* **2017**, *139* (13), 4780-4788.
157. Callegari, D.; Lodola, A.; Pala, D.; Rivara, S.; Mor, M.; Rizzi, A.; Capelli, A. M., Metadynamics Simulations Distinguish Short- and Long-Residence-Time Inhibitors of Cyclin-Dependent Kinase 8. *J. Chem. Inf. Model.* **2017**, *57* (2), 159-169.
158. Bortolato, A.; Deflorian, F.; Weiss, D. R.; Mason, J. S., Decoding the Role of Water Dynamics in Ligand-Protein Unbinding: CRF1R as a Test Case. *J. Chem. Inf. Model.* **2015**, *55* (9), 1857-1866.
159. Marchi, M.; Ballone, P., Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems. *J. Chem. Phys.* **1999**, *110* (8), 3697-3702.
160. Barducci, A.; Bussi, G.; Parrinello, M., Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100* (2), 020603.
161. Kokh, D. B.; Amaral, M.; Bomke, J.; Gradler, U.; Musil, D.; Buchstaller, H. P.; Dreyer, M. K.; Frech, M.; Lowinski, M.; Vallee, F.; Bianciotto, M.; Rak, A.; Wade, R. C., Estimation of Drug-Target Residence Times by tau-Random Acceleration Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2018**, *14* (7), 3859-3869.
162. Agius, L., Targeting hepatic glucokinase in type 2 diabetes: weighing the benefits and risks. *Diabetes* **2009**, *58* (1), 18-20.
163. Hooper, C.; Killick, R.; Lovestone, S., The GSK3 hypothesis of Alzheimer's disease. *J Neurochem* **2008**, *104* (6), 1433-9.
164. Paci, E.; Karplus, M., Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J. Mol. Biol.* **1999**, *288* (3), 441-459.
165. Petukh, M.; Stefl, S.; Alexov, E., The role of protonation states in ligand-receptor recognition and binding. *Curr. Pharm. Des.* **2013**, *19* (23), 4182-4190.
166. Spitaleri, A.; Decherchi, S.; Cavalli, A.; Rocchia, W., Fast Dynamic Docking Guided by Adaptive Electrostatic Bias: The MD-Binding Approach. *J. Chem. Theory Comput.* **2018**, *14* (3), 1727-1736.
167. Decherchi, S.; Rocchia, W., A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. *PLoS ONE* **2013**, *8* (4), e59744.
168. Berg, S.; Bergh, M.; Hellberg, S.; Hogdin, K.; Lo-Alfredsson, Y.; Soderman, P.; von Berg, S.; Weigelt, T.; Ormo, M.; Xue, Y.; Tucker, J.; Neelissen, J.; Jerling, E.; Nilsson, Y.; Bhat, R., Discovery of

- novel potent and highly selective glycogen synthase kinase-3 β (GSK3 β) inhibitors for Alzheimer's disease: design, synthesis, and characterization of pyrazines. *J. Med. Chem.* **2012**, 55 (21), 9107-9119.
169. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013**, 27 (3), 221-234.
 170. Milletti, F.; Storch, L.; Sforza, G.; Cruciani, G., New and Original pKa Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, 47 (6), 2172-2181.
 171. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, 97 (40), 10269-10280.
 172. Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A., NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comp. Phys. Comm.* **2010**, 181 (9), 1477-1489.
 173. Bas, D. C.; Rogers, D. M.; Jensen, J. H., Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **2008**, 73 (3), 765-783.
 174. Decherchi, S.; Bottegoni, G.; Spitaleri, A.; Rocchia, W.; Cavalli, A., BiKi Life Sciences: A New Suite for Molecular Dynamics and Related Methods in Drug Discovery. *J. Chem. Inf. Model.* **2018**, 58 (2), 219-224.
 175. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, 79 (2), 926-935.
 176. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, 29 (7), 845-854.
 177. Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M., PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comp. Phys. Comm.* **2009**, 180 (10), 1961-1972.
 178. Efron, B., Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, 7 (1), 1-26.
 179. Rich, R. L.; Errey, J.; Marshall, F.; Myszk, D. G., Biacore analysis with stabilized GPCRs. *Anal. Biochem.* **2011**, 409 (2), 267-272.
 180. Day, Y. S. N.; Baird, C. L.; Rich, R. L.; Myszk, D. G., Direct comparison of binding equilibrium, thermodynamic, and rate constants determined by surface- and solution-based biophysical methods. *Protein Sci.* **2002**, 11 (5), 1017-1025.
 181. Kabsch, W., Xds. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, 66 (Pt 2), 125-132.
 182. Kabsch, W., Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, 66 (Pt 2), 133-144.
 183. Evans, P., An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr. D Struct. Biol.* **2011**, 67 (4), 282-292.

184. Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S., Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Struct. Biol.* **2011**, *67* (4), 235-242.
185. McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J., Phaser crystallographic software. *J. Appl. Crystallogr.* **2007**, *40* (4), 658-674.
186. Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D., Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **2012**, *68* (Pt 4), 352-367.
187. Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K., Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (Pt 4), 486-501.
188. Murshudov, G. N.; Vagin, A. A.; Dodson, E. J., Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr. D Struct. Biol.* **1997**, *53* (3), 240-255.
189. Vagin, A. A.; Steiner, R. A.; Lebedev, A. A.; Potterton, L.; McNicholas, S.; Long, F.; Murshudov, G. N., REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D Struct. Biol.* **2004**, *60* (12 Part 1), 2184-2195.
190. Martinez, J. A.; Larion, M.; Conejo, M. S.; Porter, C. M.; Miller, B. G., Role of connecting loop I in catalysis and allosteric regulation of human glucokinase. *Protein Sci.* **2014**, *23* (7), 915-922.
191. Kamata, K.; Mitsuya, M.; Nishimura, T.; Eiki, J.; Nagata, Y., Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure* **2004**, *12* (3), 429-438.
192. Kim, Y. B.; Kalinowski, S. S.; Marcinkeviciene, J., A pre-steady state analysis of ligand binding to human glucokinase: evidence for a preexisting equilibrium. *Biochemistry* **2007**, *46* (5), 1423-1431.
193. Branduardi, D.; Gervasio, F. L.; Parrinello, M., From A to B in free energy space. *J. Chem. Phys.* **2007**, *126* (5), 054103.
194. Crooks, G. E., Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *J. Stat. Phys.* **1998**, *90* (5), 1481-1487.
195. Park, S.; Schulten, K., Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.* **2004**, *120* (13), 5946-5961.
196. Lenselink, E. B.; Louvel, J.; Forti, A. F.; van Veldhoven, J. P. D.; de Vries, H.; Mulder-Krieger, T.; McRobb, F. M.; Negri, A.; Goose, J.; Abel, R.; van Vlijmen, H. W. T.; Wang, L.; Harder, E.; Sherman, W.; Ijzerman, A. P.; Beuming, T., Predicting Binding Affinities for GPCR Ligands Using Free-Energy Perturbation. *ACS Omega* **2016**, *1* (2), 293-304.
197. Wu, Y.; Tepper, H. L.; Voth, G. A., Flexible simple point-charge water model with improved liquid-state properties. *J. Chem. Phys.* **2006**, *124* (2), 024503.
198. Berendsen, H.; Grigera, J.; Straatsma, T., The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91* (24), 6269-6271.

199. Matos, G. D. R.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; Mobley, D. L., Approaches for calculating solvation free energies and enthalpies demonstrated with an update of the FreeSolv database. *Journal of Chemical & Engineering Data* **2017**, 62 (5), 1559-1569.
200. Blondel, A.; Karplus, M., New formulation for derivatives of torsion angles and improper torsion angles in molecular mechanics: Elimination of singularities. *J. Comput. Chem.* **1996**, 17 (9), 1132-1141.
201. Grønbech-Jensen, N.; Farago, O., A simple and effective Verlet-type algorithm for simulating Langevin dynamics. *Mol. Phys.* **2013**, 111 (8), 983-991.
202. Martins, S. A.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A., Prediction of Solvation Free Energies with Thermodynamic Integration Using the General Amber Force Field. *J. Chem. Theory Comput.* **2014**, 10 (8), 3570-7.
203. Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J., The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput. Aided Mol. Des.* **2010**, 24 (4), 259-279.
204. Birch, F., Finite Elastic Strain of Cubic Crystals. *Phys. Rev.* **1947**, 71 (11), 809-824.
205. Stillinger, F. H.; Weber, T. A., Inherent structure in water. *The Journal of Physical Chemistry* **1983**, 87 (15), 2833-2840.
206. Soper, A. K., The Radial Distribution Functions of Water as Derived from Radiation Total Scattering Experiments: Is There Anything We Can Say for Sure? *ISRN Physical Chemistry* **2013**, 2013, 1-67.
207. Soper, A. K., The radial distribution functions of water and ice from 220 to 673 K and at pressures up to 400 MPa. *Chem. Phys.* **2000**, 258 (2), 121-137.
208. Andanson, J. M.; Traïkia, M.; Husson, P., Ionic association and interactions in aqueous methylsulfate alkyl-imidazolium-based ionic liquids. *J. Chem. Thermodyn.* **2014**, 77, 214-221.
209. Rastogi, A.; Ghosh, A. K.; Suresh, S., Hydrogen bond interactions between water molecules in bulk liquid, near electrode surfaces and around ions. In *Thermodynamics-Physical Chemistry of Aqueous Systems*, InTech: 2011.
210. Maréchal, Y., The molecular structure of liquid water delivered by absorption spectroscopy in the whole IR region completed with thermodynamics data. *J. Mol. Struct.* **2011**, 1004 (1), 146-155.
211. Pauling, L., The Structure and Entropy of Ice and of Other Crystals with Some Randomness of Atomic Arrangement. *J. Am. Chem. Soc.* **1935**, 57 (12), 2680-2684.
212. Bernal, J. D.; Fowler, R. H., A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions. *J. Chem. Phys.* **1933**, 1 (8), 515-548.
213. Vega, C.; Sanz, E.; Abascal, J. L. F., The melting temperature of the most common models of water. *J. Chem. Phys.* **2005**, 122 (11), 114507.
214. Burkard, R.; Dell'Amico, M.; Martello, S., *Assignment Problems*. Society for Industrial and Applied Mathematics: 2009; p 402.
215. Berryman, J. T.; Schilling, T., Free Energies by Thermodynamic Integration Relative to an Exact Solution, Used to Find the Handedness-Switching Salt Concentration for DNA. *J. Chem. Theory Comput.* **2013**, 9 (1), 679-686.

216. Kuhn, H. W., The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **1955**, 2 (1-2), 83-97.
217. Wagner, W.; Pruß, A., The IAPWS Formulation 1995 for the Thermodynamic Properties of Ordinary Water Substance for General and Scientific Use. *J. Phys. Chem. Ref. Data* **2002**, 31 (2), 387-535.
218. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H., PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016**, 44 (D1), D1202-13.
219. Briard, P.; Rossi, J. C., Ketoprofene. *Acta Crystallogr. C* **1990**, 46 (6), 1036-1038.
220. Serjeant, E. P.; Dempsey, B., Ionisation constants of organic acids in aqueous solution. *Pergamon Press, Oxford; New York* **1979**.
221. Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M., DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, 46 (Database issue), D1074-D1082.
222. <https://www.drugbank.ca/drugs/DB01009>.
223. Grimme, S., Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, 27 (15), 1787-1799.
224. Troullier, N.; Martins, J. L., Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B* **1991**, 43 (3), 1993.
225. Stillinger Jr, F. H.; Lovett, R., General restriction on the distribution of ions in electrolytes. *J. Chem. Phys.* **1968**, 49 (5), 1991-1994.
226. Witherspoon, P. A.; Saraf, D. N., Diffusion of Methane, Ethane, Propane, and n-Butane in Water from 25 to 43°. *The Journal of Physical Chemistry* **1965**, 69 (11), 3752-3755.
227. Freitas, R.; Asta, M.; de Koning, M., Nonequilibrium free-energy calculation of solids using LAMMPS. *Comput. Mater. Sci.* **2016**, 112, 333-341.
228. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, 26 (16), 1668-1688.

Abstract

Virtually all biochemical activities are mediated by the organization and recognition of biological macromolecules. An accurate characterization of the thermodynamics and kinetics governing the formation of supramolecular complexes is required to deeply understand the molecular principles driving all biological interactions. Thermodynamics provides the driving force of protein-ligand binding and is quantified by the binding free energies or the equilibrium dissociation constants. Since the interacting partners are out of equilibrium in vivo, the thermodynamic description of binding needs to be complemented by the knowledge of the kinetic rates. Nowadays, various biophysical experimental techniques can determine thermodynamic and kinetic properties, which are still difficult to be efficiently predicted by computational methods mainly because of the limited force field accuracy and the high computational cost.

During my Ph.D., I applied molecular dynamics (MD)-based methods to characterize the thermodynamics and kinetics of inter-molecular interactions. First, I worked on a new enhanced MD-based protocol to simulate protein-ligand dissociation events. This approach provides a realistic description of the evolution of the system to an external perturbation accounting for the natural forces driving the dissociation mechanisms. By applying this computational approach to two pharmaceutically relevant kinases, I was able to rank two series of compounds on unbinding kinetics and to get qualitative mechanistic and path information on the underlying unbinding events, providing additional valuable information to be used in the optimization of lead compounds. Then, I developed an innovative computational method to estimate free energies applicable to systems of arbitrary complexity. Despite the number of challenges to be overcome, the method is very promising being able to provide accurate free energy estimates. Therefore, computer simulations emerged as a valuable tool to obtain information on both the thermodynamic and kinetic aspects governing the formation of supramolecular complexes, which might be used to assist the early phases of the drug discovery pipeline.